Progressive and biased divergent evolution underpins the origin and diversification of peridinin dinoflagellate plastids

Dorrell, Richard G.; Klinger, Christen M.; Newby, Robert J.; Butterfield, Erin R.; Richardson, Elisabeth; Dacks, Joel B.

Published in:
Molecular Biology and Evolution

DOI:
10.1093/molbev/msw235

Publication date:
2016

MBE Advance Access published November 4, 2016

Articles: Discoveries

Progressive and biased divergent evolution underpins the origin and diversification of peridinin dinoflagellate plastids

Richard G. Dorrell[1+], Christen M. Klinger*[2], Robert J. Newby*[3], Erin R. Butterfield[4, 5], Elisabeth Richardson[2], Joel B. Dacks[2], Christopher J. Howe[6], R. Ellen R. Nisbet[6], and Chris Bowler[1]

1 Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France
2 Department of Cell Biology, University of Alberta
3 Department of Biology, Middle Tennessee State University
4 Department of Biochemistry, Pennsylvania State University
5 School of Life Sciences, University of Dundee
6 Department of Biochemistry, University of Cambridge
* Contributed equally to this manuscript
+ To whom correspondence should be addressed: dorrell@biologie.ens.fr

Abstract

Dinoflagellates are algae of tremendous importance to ecosystems and to public health. The cell biology and genome organisation of dinoflagellate species is highly unusual. For example, the plastid genomes of peridinin-containing dinoflagellates encode only a minimal number of genes arranged on small elements termed "minicircles". Previous studies of peridinin plastid genes have found evidence for divergent sequence evolution, including extensive substitutions, novel insertions and deletions, and use of alternative translation initiation codons. Understanding the extent of this divergent evolution has been hampered by the lack of characterised peridinin plastid sequences. We have identified over 300 previously unannotated peridinin plastid mRNAs from published transcriptome projects, vastly increasing the number of sequences available. Using these data, we have produced a well-resolved phylogeny of peridinin plastid lineages, which uncovers several novel relationships within the dinoflagellates. This enables us to define changes to plastid sequences that occurred early in dinoflagellate evolution, and that have contributed to the subsequent diversification of individual dinoflagellate clades. We find that the origin of the peridinin dinoflagellates was specifically accompanied by elevations both in the overall number of substitutions that occurred on plastid sequences, and in the Ka/Ks ratio associated with plastid sequences, consistent with changes in selective pressure. These substitutions, alongside other changes, have accumulated progressively in individual peridinin plastid lineages. Throughout our entire dataset, we identify a persistent bias towards non-synonymous substitutions occurring on sequences encoding photosystem I subunits and stromal regions of peridinin plastid proteins, which may have underpinned the evolution of this unusual organelle.

Introduction

Dinoflagellates are eukaryotes with immense ecological and evolutionary importance. They include photosynthetic, heterotrophic, mixotrophic, and parasitic representatives (Dorrell and Howe 2015). The photosynthetic species are a major component of plankton communities in marine and freshwater environments (Leterme, et al. ; de Vargas, et al. 2015), and include causative agents of harmful algal blooms (*Protoperidinium, Ceratium*) (Hallegraeff 2010; Hinder, et al. 2012), symbionts of corals (*Symbiodinium*) and marine protozoa (*Pelagodinium, Brandtodinium*) (Siano, et al. 2010; Probert, et al. 2014). The non-photosynthetic species include parasites of marine invertebrates (*Hematodinium,* Syndiniales, and ellobiopsids) (Gornik, et al. 2012), and bioluminescent phagotrophs (*Noctiluca*) (Nakamura 1998). Dinoflagellates are members of the alveolates, a group that additionally contains important laboratory model species (ciliates such as *Paramecium* and *Tetrahymena*), pathogens of terrestrial and marine animals (e.g., the apicomplexan parasites *Plasmodium* and *Toxoplasma*, and the mollusc pathogen *Perkinsus*), ecologically significant heterotrophs (ciliates and gregarines). Two other photosynthetic alveolates have been described (the "chromerids" *Chromera* and *Vitrella*), both of which are closely related to the apicomplexans, and serve as model organisms for understanding the origins of parasitism in this lineage (Janouskovec, et al. 2010; Dorrell and Howe 2015).

Dinoflagellates are renowned for their unusual and distinctive cell biology (Lin 2011; Wisecaver and Hackett 2011; Dorrell and Howe 2015). Unlike those of all other studied eukaryotes, dinoflagellate chromosomes are maintained in a permanently condensed state, and do not principally utilise histones for DNA packaging (Gornik, et al. 2012). The mitochondrial genomes of dinoflagellates and apicomplexans are highly reduced, and several otherwise well-conserved subunits of the ATP synthase complex have not been documented in any dinoflagellates (Butterfield, et al. 2013; Janouškovec, et al. 2013).

One of the oddest traits of dinoflagellates is their possession of unusual plastids. While all other major eukaryotic groups are either non-photosynthetic, or contain only one plastid lineage, at least four and potentially as many as seven phylogenetically distinct plastid lineages have been documented across the dinoflagellates (Janouskovec, et al. 2010; Dorrell and Howe 2015). The majority of photosynthetic dinoflagellates possess plastids that contain the light harvesting pigment peridinin (Haxo, et al. 1976). These "peridinin plastids" are most likely of red algal origin, although the exact endosymbiotic events through which they originated remain debated (Keeling 2010; Ševčíková, et al. 2015). Phylogenetic studies have indicated that the peridinin plastid was present in a common ancestor of dinoflagellates, chromerids, and apicomplexans, and that the alternative plastids found in some dinoflagellates have originated through serial endosymbiosis (Janouskovec, et al. 2010; Dorrell and Howe 2015).

The genome of the peridinin plastid is the smallest known from a photosynthetic plastid, retaining only twelve protein-coding genes (Barbrook, et al. 2014; Mungpakdee, et al. 2014), plus ribosomal RNA and some transfer RNA genes (Barbrook, et al. 2006; Nelson, et al. 2007). The protein coding genes solely encode core subunits of the photosynthetic electron transport machinery, comprising genes encoding six subunits of photosystem II (*psbA, psbB, psbC, psbD, psbE, psbI*), and two subunits each of photosystem I (*psaA, psaB*), cytochrome b_6/f (*petB, petD*) and plastid ATP synthase (*atpA, atpB*). These genes are located on small elements termed "minicircles" (of 1600-6600 bp length) (Zhang, et al. 1999; Nelson and Green 2005) and "microcircles" (of 400- 600bp length) (Nisbet, et al. 2004). Minicircles

typically contain single genes, although minicircles that contain no genes (Nisbet, et al. 2004; Nelson, et al. 2007), or combinations of multiple genes have also been identified in multiple dinoflagellate species (Hiller 2001; Nisbet, et al. 2004; Nelson, et al. 2007). All other genes of plastid origin that have been documented in peridinin dinoflagellates are located in the nucleus (Morse, et al. 1995; Bachvaroff, et al. 2004; Mungpakdee, et al. 2014). Some minicircles contain sizeable open reading frames of unknown function that are unique to dinoflagellates (Barbrook, et al. 2001; Nisbet, et al. 2004; Barbrook, et al. 2006). It has been proposed that some peridinin plastids contain genes encoding ribosomal proteins (*rpl28, rpl33*) and iron-sulfur cluster biogenesis factors (*ycf16, ycf24*), which were acquired by horizontal gene transfer from non-photosynthetic bacteria, although whether these sequences are genuinely plastid-encoded remains controversial (Moszczynski, et al. 2012; Dorrell and Howe 2015).

Alongside the extreme level of reduction observed in the peridinin plastid genome, the transcript sequences produced in peridinin plastids are highly unusual. This is in part due to the unusual transcript processing machinery associated with peridinin plastids, which includes (in some species) extensive in-frame sequence editing (Zauner, et al. 2004; Dorrell and Howe 2015) and (in all documented species) the addition of a 3' poly(U) tail to plastid mRNAs (Wang and Morse 2006; Barbrook, et al. 2012), which contrast with the 3' poly(A) tail and 5' spliced-leader sequences added to transcripts in dinoflagellate nuclei (Zhang, et al. 2007). In addition, individual genes within peridinin plastids are highly divergent from orthologues from other plastid lineages (Shalchian-Tabrizi, et al. 2006; Pochon, et al. 2014). Genes encoded located in peridinin plastids frequently contain extensive sequence substitutions (Barbrook, et al. 2014) and in-frame insertions and deletions (Barbrook, et al. 2006; Barbrook, et al. 2014), have unusual codon usage preferences (Inagaki, et al. 2004; Bachvaroff, et al. 2006) and use a range of alternative translation initiation codons in addition to ATG (ATA, ATT, GTA, TTG) (Zhang, et al. 1999; Barbrook, et al. 2014).

This project was conceived to investigate the timing and extent of the divergent sequence evolution in peridinin plastids. It is broadly agreed that application of poly(U) tails to plastid transcripts is an ancestral feature of peridinin dinoflagellates, and it has been proposed that the fragmentation of the peridinin plastid genome into minicircles, and the reduction of the plastid genome to a minimal protein-coding gene set, are likewise ancestral (Janouskovec, et al. 2010; Dorrell, et al. 2014; Dorrell and Howe 2015). However, it is not known when other divergent evolutionary events occurred in peridinin plastids. This has in part been due to the lack of available sequence information, with essentially complete plastid coding sequences previously available only for *Amphidinium carterae* (Nisbet, et al. 2004), and for two strains of *Symbiodinium* (Clade C3, and Clade Mf) (Barbrook, et al. 2014; Mungpakdee, et al. 2014). In addition, the phylogenetic relationships within the peridinin dinoflagellates remain poorly resolved. While *Amphidinium* is agreed to diverge at the base of the peridinin dinoflagellates, the branching order of other lineages remains debated (Hoppenrath and Leander 2010; Bachvaroff, et al. 2014; Gavelis, et al. 2015). We wished to generate a robust phylogeny of extant peridinin dinoflagellate lineages, and use this phylogeny to answer three questions: (1) which of the divergent features associated with peridinin plastid sequences probably have arisen in an ancestor to all species studied; (2) to what extent and in which dinoflagellate lineages have divergent plastid evolution events occurred more recently, and (3) whether there are any consistent trends across the dinoflagellates in terms of which plastid-encoded proteins, or regions of plastid-encoded proteins, are the most divergent, extending from their common ancestor through to extant species.

We present a taxonomically detailed reconstruction of the evolution of peridinin plastid sequences. We have focused on identifying novel plastid transcripts, which allows us to assess the composite effects of divergent gene evolution and transcript editing on peridinin plastid sequences (Zauner, et al. 2004; Dorrell and Howe 2015). We have annotated over 300 new peridinin plastid sequences from published transcriptome resources, and demonstrate that the twelve photosystem genes previously identified in peridinin plastids (Mungpakdee, et al. 2014; Dorrell and Howe 2015) probably represent the complete protein-coding component of the plastid genome of a common ancestor of all extant photosynthetic dinoflagellates. We have used these sequences to generate a well-resolved phylogeny of peridinin plastids, uncovering novel evolutionary relationships between among the major dinoflagellate lineages, and have used this phylogeny to determine when different divergent evolutionary events are have occurred in peridinin plastids. To disentangle the different factors that have underpinned this unusual sequence evolution, we have calculated substitution rates, and Ka/Ks ratios (also referred to as dN/dS, and defined as the relative enrichment in non-synonymous substitutions, Ka, to synonymous substitutions, Ks, over a particular sequence, which provides an indicator of the strength of selective pressure (Yang and Bielawski 2000; Hurst 2002)) for each dinoflagellate species and each residue of each plastid-encoded protein studied.

We show that the origin of the peridinin plastid was specifically marked by an elevation in substitution rates and in Ka/Ks ratios, consistent with changes in selection pressure in the dinoflagellate common ancestor. We additionally show that these and other divergent features, such as in-frame insertions and alternative translation initiation codons, have continued to evolve progressively in individual dinoflagellate clades. Finally, we show that in both the common ancestor of all studied peridinin dinoflagellates and in extant species, elevated Ka/Ks ratios are concentrated on genes encoding photosystem I subunits, and codons encoding stromal-facing residues of plastid proteins. This may suggest that specific changes to dinoflagellate physiology have driven the divergent evolution of the peridinin plastid. Ultimately, our study provides valuable insights into the evolutionary history of this unusual plastid lineage, from its very origins to subsequent diversification.

Results

1. Annotation of new peridinin plastid sequences from transcriptome data

Identification of plastid sequences

New dinoflagellate orthologues of the twelve protein-coding genes (*atpA, atpB, petB, petD, psaA, psaB, psbA, psbB, psbC, psbD, psbE* and *psbI*) previously found to be retained in peridinin plastids (Howe, et al. 2008; Barbrook, et al. 2014; Mungpakdee, et al. 2014) were identified in the NCBI EST library, and transcriptome libraries within the Marine Microeukaryote Transcriptome Sequencing Project (MMETSP) (Keeling, et al. 2014). A total of 381 sequences were identified, 364 of which had no previously annotated equivalents (Fig. 1; Table S1). The majority (348) of the novel sequences were identified from previously unannotated transcripts within the MMETSP libraries, with only a few sequences (16) identifiable from EST libraries located in NCBI (Fig. 1). Sequences of plausible plastid origin were identified in all but a few peridinin dinoflagellate MMETSP libraries, with a minimum of 15 novel orthologues found for each gene studied (Table S1), and complete sets of protein-coding plastid genes identified for 13 new dinoflagellate species in addition to the three already characterised (Howe, et al. 2008; Barbrook, et al. 2014; Mungpakdee, et al. 2014).

Novel sequences are of probable plastid origin

Uninterrupted poly(T) stretches of 4 bp or longer were detected on the 3' end of half (191/381) of the novel dinoflagellate sequences, consistent with the presence of the poly(U) tails associated with dinoflagellate plastid transcripts (Fig. S1; Table S1). Across the entire dataset, only one sequence was found that terminated in a possible 3' poly(A) tail, and none contained evidence of 5' spliced-leader sequences, or plausible tripartite targeting sequences (consisting of a signal peptide, ASAFAP-delimited transit peptide, and a downstream hydrophobic region), as are associated with nucleus-encoded, plastid-targeted proteins in peridinin dinoflagellates (Nassoury and Morse 2005; Zhang, et al. 2007) (Table S1).

Limited additional coding sequences in peridinin plastids

The entire assembled transcriptome dataset was screened for further transcripts that might originate from peridinin plastids. We could not find convincing evidence for any further genes that are plastid-encoded in other non-dinoflagellate lineages, and might still be plastid-encoded within individual dinoflagellates, beyond the twelve photosystem genes previously documented (Fig. S2; Supplementary Results, Section 1). The entire dataset was additionally searched for homologues of the four plastid genes (*rpl28, rpl33, ycf16, ycf24*) proposed to have been gained by horizontal transfer from bacteria into specific peridinin plastids (Moszczynski, et al. 2012). A total of 148 new sequences were identified and inspected using single-gene phylogenies. While the original *Pyrocystis lunula* and *Ceratium horridum* sequences grouped with Bacteroidetes, none of the homologues within this dataset did: instead, the majority resolved as a monophyletic group and all grouped either within, or as sister-groups to, other plastid or cyanobacterial lineages, with robust (>90%) bootstrap support (Fig. S3-S6). None of the *rpl28, rpl33, ycf16* or *ycf24* transcript sequences identified in this study possessed a 3' poly(U) tail (Table S1); however, many contained 3' poly(A) tails, spliced-leader sequences (Table S1), and tripartite plastid targeting sequences (Fig. S7), consistent with a nuclear origin.

Homologues of the four novel open reading frames previously identified on minicircles located in the *Amphidinium carterae* plastid (Barbrook and Howe 2000; Dorrell and Howe 2015) were searched for across the entire transcript dataset. Only equivalents of *A. carterae ORF1, ORF2* and *ORF3* were detected, and these were limited to the related species *Amphidinium massartii* (Fig. S8, panel A). The *ORF*-like sequences identified in *A. massartii* were highly divergent from those of *A. carterae*, with only 46% (*ORF1)*, 32% (*ORF2*) and 33% (*ORF3*) between the two sequences, compared to (for example) 97% sequence conservation between the *psbD* sequences from each species (Fig. S8, panel A). We could find only limited evidence for the presence of additional conserved polyuridylylated transcripts that might correspond to novel plastid ORFs within the dataset (Fig. S8; Supplementary Results, Section 1).

2. Reconstruction of phylogenetic relationships between peridinin dinoflagellates

A concatenated protein alignment (3,410 amino acids, average 72.7% pairwise identities) was generated, consisting of the twelve plastid sequences studied, for each of the dinoflagellates present in MMETSP, and a reference set of fifteen non-dinoflagellates (Table S3). Bayesian and Maximum Likelihood trees were generated from this alignment (Fig. 2; Table S4). Two phylogenetically distinct sets of plastid sequences were identified for clade A *Symbiodinium*, one of which resolved with other *Symbiodinium* species, and the other as a

sister group to *A. carterae*, which presumably represents a contamination within the *Symbiodinium* A MMETSP library (Fig. 2; Table S1).

Consistent with previous data, *Amphidinium* was identified as the earliest diverging peridinin dinoflagellate genus (Fig. 2) (Dorrell and Howe 2015; Gavelis, et al. 2015). Following *Amphidinium,* the next most basal dinoflagellates were *Togula jolla*, and a clade consisting of Prorocentrales, Peridiniales, and the (previously Suessialean) species *Pelagodinium beii*, which diverged from a clade consisting of Gonyaulacales, Suessiales, and the (previously Peridinialean) genus *Heterocapsa*, with moderate to robust support (in Bayesian analysis, >60% in ML trees) (Fig. 2). Identical (Fig. S9) or nearly identical (Fig. S10) topologies to the original tree were obtained in trees calculated from alignments from which long branches, fast evolving sites, or individual plastid genes had been removed (Supplementary Results, Section 2), suggesting that the initial tree topology was largely accurate.

3. Dinoflagellate-wide changes in plastid sequence composition

Changes in plastid GC content may have occurred before the radiation of the dinoflagellates

The first, second and third position GC content were compared across a 4,478 nt gap-free alignment of six plastid genes (*psaA, psaB, psbA, psbB, psbC, psbD*), for each of the dinoflagellates studied, and all of the non-dinoflagellate sequences previously used for the multigene phylogeny (Fig. 2). Elevated GC contents were observed at third codon positions in many of the dinoflagellates compared to non-dinoflagellate species (Fig. S11). These included third position GC content values of > 35% in three of the earliest diverging dinoflagellate clades (Fig. S11; *Amphidinium* GC[3]= 39.6-49.6%, Peridiniales GC[3]= 29.8-40.9%, Prorocentrales GC[3]= 31.7-39.5%). However, the chromerid *Vitrella brassicaformis*, which forms the closest sister-group to the dinoflagellates within the multigene tree (Fig. 2), also has a high third position GC content (Fig. S11; 46.9%), so it is possible that this GC enrichment is not specific to dinoflagellates.

Limited changes to plastid translation in the dinoflagellate ancestor

Across the entire 4,478 nt gap-free plastid alignment, 18 codons occurred with lower frequency (and 19 codons with higher frequency) in dinoflagellates compared to non-dinoflagellate species (one-way ANOVA, $P < 0.05$; Table S5; Fig. S12). Ten amino acids were likewise found to occur with lower frequency, and 5 with higher frequency, in dinoflagellates (one-way ANOVA, $P < 0.05$; Table S5; Fig. S12) over this alignment, as inferred by a standard translation table. We could not identify any convincing evidence for changes to plastid translation tables within the dinoflagellates (Table S6; Fig. S13; Supplementary Results, Section 3).

A much smaller number of codons were found to have undergone specific de-enrichments (6) or enrichments (8; chi-squared test, $P< 0.05$; Table S5, Fig. S12) in a common ancestor of all studied dinoflagellates (inferred by comparing the regressed ancestral sequence of all dinoflagellates studied to that of the regressed ancestral sequence of the common ancestor of all studied dinoflagellates and *Vitrella brassicaformis*, which was the closest sister-group to dinoflagellates included in the alignment (Janouskovec, et al. 2010)). Comparing the two datasets, only four codons (ATA-Ile, AGA-Arg, ACC-Thr, and TAT-Tyr) were found both to occur at significantly higher frequencies in dinoflagellates than non-dinoflagellates, and to have undergone a specific enrichment in the common ancestor of all studied dinoflagellates (Table S5; Fig. S12). Similarly, only one codon (CGT-Arg) was found to occur at a significantly

lower frequency in dinoflagellates, and to have undergone a specific de-enrichment in the common ancestor of all studied dinoflagellates (Table S5; Fig. S12). Finally, only one amino acid (Tyrosine) was found to have undergone a significant change in frequency in the common ancestor of all studied dinoflagellates (Table S5; Fig. S12), suggesting overall that relatively limited changes to plastid codon usage are associated with dinoflagellate origins.

4. Dinoflagellate-wide changes to plastid sequence evolution

Elevated pairwise substitutions at the origin of dinoflagellates

Total numbers of pairwise substitutions were calculated for every species used in the multigene phylogeny, over the 4,478 nucleotide gap-free alignment (Fig. 3; Table S7). The dinoflagellate sequences were highly divergent from the non-dinoflagellate species (Fig. 3; compare top left hand corner of figure to remainder). On average, dinoflagellate and non-dinoflagellate species pairs were separated by 1,549 nucleotide substitutions, while pairs of non-dinoflagellate species were separated by an average of 994 substitutions, which was significantly lower (Table S7; Fig. S14, panel A; one-way ANOVA, $P= 1.86 \times 10^{-179}$). Even the related alveolate lineage *Vitrella brassicaformis*, which was separated from other non-dinoflagellate species by a much larger number of substitutions (average value 1276) was significantly less divergent than dinoflagellate species studied (Table S7; Fig. S14, panel A; $P= 1.2 \times 10^{-18}$), suggesting that this elevated substitution rate is specifically associated with dinoflagellate species.

We tested whether the elevated numbers of substitutions observed between dinoflagellate and non-dinoflagellate species were related either to plastid GC content, codon usage, or amino acid composition (Figs. S14, S15; Supplementary Results, Section 4). While changes to plastid GC content and codon usage were correlated to the total numbers of pairwise substitutions observed (Fig. S15), alignment recoding to remove these effects did not eliminate the elevated substitution rates associated with dinoflagellate species (Fig. S14, panel B).

Elevated Ka/Ks ratios at the origin of dinoflagellates

Pairwise Ka/Ks ratios (which provide information regarding the strength of selective pressure acting on individual sequences, as expressed by the ratio of non-synonymous to synonymous substitutions) were additionally calculated for each species pair (Fig. 3; Table S7). Similar to the situation observed for total substitution rates, much higher pairwise Ka/Ks ratios were observed between dinoflagellate and non-dinoflagellate species pairs (average value 0.0592) than within non-dinoflagellate species pairs (Fig. S16, panel A; average value 0.0183; $P= 2.75 \times 10^{-111}$). A dramatic difference was also observed for the Ka/Ks ratios calculated between dinoflagellates and non-dinoflagellate species, and *Vitrella* and all other non-dinoflagellate species (Fig. S16, panel A; average value 0.0271; $P= 2.87 \times 10^{-18}$).

We tested whether the elevated Ka/Ks ratios observed in dinoflagellates might be related to the extremely high total numbers of pairwise substitutions observed, for example due to a saturating substitution rate at codon third positions leading to an underestimate of the total synonymous substitutions between dinoflagellate and non-dinoflagellate species (Figs. S17-S18; Supplementary Results, Section 5). The third position substitution rates between dinoflagellate and non-dinoflagellate species, while high, were not at saturation rate (Figs. S17, S18; Supplementary Results, Section 5). Other variables tested were either not correlated to the pairwise Ka/Ks ratios obtained (in the case of amino acid composition), or

were not sufficient (judged by alignment recoding) to explain the differences in Ka/Ks ratios observed (in the case of third position GC content, and codon usage; Figs. S16, S19; Supplementary Results, Section 6).

5. Lineage-specific changes to peridinin plastid sequences

Extremely elevated Ka/Ks ratios within dinoflagellates

Some dinoflagellate species were found to have extremely elevated pairwise Ka/Ks ratios calculated relative to other dinoflagellates in the alignment (Fig. 3). For example, *Pelagodinium beii* was separated from all other dinoflagellates by a minimum number of 1,400 substitutions, and *Brandtodinium nutriculum* was separate from all other dinoflagellates by a minimum number of 1,399 substitutions (Fig. 3; Table S7), both of which were far larger than the average minimum number of total pairwise substitutions (416) calculated for other dinoflagellate species (Z test, $P < 0.05$). These separations were found to be independent of plastid GC content and codon usage patterns in both species (Fig. S20; Supplementary Results, Section 8). Both species were also found to have elevated Ka/Ks ratios (minimum *P. beii* Ka/Ks 0.0618; minimum *B. nutriculum* Ka/Ks 0.0408; average dinoflagellate minimum Ka/Ks 0.0174; Fig. 3), but these differences were eliminated by removing codon third positions from Ka/Ks calculations, suggesting that they are the result of saturating mutation rates in each species (Fig. S20; Supplementary Results, Section 8).

A dramatic evolutionary divergence was observed within members of the Gonyaulacales and Suessiales (Fig. 3, bottom right hand sector; Fig. S21). Species within these lineages had extremely high pairwise Ka/Ks ratios, with a maximum value of 0.284 between the Gonyaulacales *Protoceratium reticulatum* and *Pyrodinium bahamense*, which is significantly greater than the average maximum pairwise Ka/Ks ratio (0.140) observed for all dinoflagellates (Fig. 3; Table S7; Z test, $P< 0.05$). The average pairwise Ka/Ks ratio between Gonyaulacalean and Suessialean dinoflagellates (0.110) was significantly higher than the average pairwise Ka/Ks ratio between other dinoflagellate pairs (0.043; Fig. S21, panel A; $P=1.57 \times 10^{-09}$). This was found to be independent of both third position substitutions, and changes to codon usage in these species (Supplementary Results, Section 8). In contrast, the average number of pairwise substitutions between Gonyaulacalean and Suessialean species (711) was significantly lower ($P=0$) than the average pairwise substitution frequencies observed between other dinoflagellates (1417; Fig. 3; Fig. S21, panel B), indicating that the rapid divergent evolution within this lineage is specifically due to an elevated Ka/Ks ratio.

Widespread evolution of alternative translation initiation codons in peridinin plastids

44% of all the sequences identified lacked a plausible ATG initiation codon, hence probably use alternative translation initiation codons (Table S8). Eight different codons were identified as probable alternative initiation sites for individual peridinin plastid transcripts, with TTG and ATT occurring the most frequently (Fig.4; Table S8). None of the alternative translation initiation codons identified was conserved across all dinoflagellates, and the majority were species-specific (Fig. 4, panel A). However, twenty alternative initiation codons were conserved across multiple dinoflagellate species (Fig. 4, panel B, square labels), such as the adoption of a TTG alternative initiation codon in *psbA* in a common ancestor of *Alexandrium, Gambierdiscus, Pyrodinium* and *Gonyaulax sp.* (Fig. 4, panel B, label I; Fig. S22). Some lineages utilise alternative initiation codons more frequently than others, with six alternative initiation codons (petB-GTG, petD- KTG, psaB-ATW, psbC-ATC, psbD-ATY, psbE-

ATT) found in *Scrippsiella hangoei* and its sister species *Peridinium aciculiferum* (Fig. 4, panel B, labels C, D), compared to only one (atpA-TTG) in *Amphidinium* sp. (Fig. 4, panel B, label A).

Multiple discrete changes to plastid protein sequences within dinoflagellates

A total of 111 insertions and 48 deletions distributed across 80 positions were identified within dinoflagellates (Fig. 4, Panel A; Table S9). This contrasts to the situation for the non-dinoflagellate species in the alignment, for which only 19 insertions and 22 deletions were identified (Table S9). Twelve insertions and five deletions were conserved across all dinoflagellates (Fig. 4, panel A; Fig. S23, panel A), although many of these indels have undergone substantial expansions or contractions in individual species (Fig. S23, panel B). The majority of indels, however, were restricted to individual dinoflagellate species (77 insertions, 25 deletions) or clades (22 insertions, 18 deletions) (Fig. 4, panel A; Fig. S23, panel C). These included two insertions (in PsaB, insertion starting at consensus residue 41; and PsbC, residue 197) and one deletion (in PsbB, starting at consensus residue 291) that evolved in a common ancestor of all species studied, except for the basally divergent *Amphidinium*, and one insertion (PsaB, residue 168) and one deletion (PsaB, residue 602) in a common ancestor of Gonyaulacales and Suessiales (Fig. 4, panel B; triangular labels).

Instances in which a residue that is conserved in all other plastid species studied was lost in dinoflagellates were tabulated for one representative protein, the ATP synthase subunit AtpA. 200 such changes were found, of which only 15 were ancestral (Fig. 4, panel A; Table S10). Of the remaining 185 substitutions, 93 were species-specific, while 92 were shared across specific dinoflagellate clades. These clade-specific changes included the loss of thirteen otherwise conserved residues in a common ancestor of all genera except *Amphidinium,* and two in a common ancestor of Gonyaulacales and Suessiales (Fig. 4, panel B, circular labels; Table S10).

6. Identification of consistent trends in peridinin plastid evolution

Photosystem I sequences in peridinin plastids have elevated Ka/Ks ratios

An elevated dinoflagellate Ka/Ks ratio was observed in the two photosystem I subunit genes (*psaA* and *psaB*) retained in dinoflagellate plastids. For *psaA*, the dinoflagellate Ka/Ks (0.481) was 4.89 times that of the non-dinoflagellate value (0.098); whereas for *psaB* the dinoflagellate Ka/Ks (0.446) was 4.47 times that of the non-dinoflagellate value (0.099; Fig. 5, panel A). Both the Ka/Ks ratios observed were substantially greater than those observed for dinoflagellate sequences across the entire dataset, both in terms of the raw Ka/Ks ratio (0.296) and in terms of the relative enrichment (2.41) compared to the non-dinoflagellate Ka/Ks ratio (0.123; Fig. 5, panel A). Elevated Ka/Ks values were also obtained for dinoflagellate *psaA* and *psaB* genes in alignments recoded to eliminate third position substitutions and codon usage bias (Fig. S24; Tables S11, S12-S15; Supplementary Results, Section 8).

Photosystem I sequences in the dinoflagellate ancestor also had elevated Ka/Ks ratios

Individual Ka/Ks ratios were also calculated for each sequence, solely between the hypothetical sequences calculated for the common ancestor of all studied dinoflagellates, and the common ancestor of dinoflagellates and *Vitrella*, which should correspond to the substitutions that most probably occurred in dinoflagellates immediately following their divergence from other plastid lineages (Table S11). Elevated Ka/Ks ratios were found in both

the dinoflagellate ancestor *psaA* (ancestral Ka/Ks 0.425; dinoflagellate ancestor/non-dinoflagellate ratio 4.32) and *psaB* sequences (ancestral Ka/Ks 0.593; dinoflagellate ancestor/non-dinoflagellate ratio 5.95), compared to all other genes (ancestral Ka/Ks 0.376; dinoflagellate ancestor/non-dinoflagellate ratio 3.07; Fig. 5, panel A; Tables S11, S12). The Ka/Ks ratios associated with the *psaB* dinoflagellate ancestor sequence were confirmed using alignment recoding to be genuine, rather than a consequence of a highly elevated mutation rate or change in codon usage preference specific to dinoflagellate photosystem I genes (Fig. S24; Supplementary Results, Section 8).

Elevated Ka/Ks ratios are specific to photosystem I genes

Only one other gene, *psbI*, was found to have a greater than average Ka/Ks ratio between dinoflagellates and non-dinoflagellates (dinoflagellate Ka/Ks 0.451; non-dinoflagellate Ka/Ks 0.132; dinoflagellate/non-dinoflagellate ratio 3.39) and between the dinoflagellate ancestor and non-dinoflagellates (dinoflagellate ancestor Ka/Ks 0.512; dinoflagellate ancestor/non-dinoflagellate ratio 3.84; Fig. 5, panel A; Table S11). However, the elevated *psbI* Ka/Ks ratio was eliminated in alignments recoded to eliminate third position substitutions (Fig. S24; Supplementary Results, Section 8), suggesting that it is the result of due a saturating mutation rate at dinoflagellate *psbI* third codon positions. No other genes (except for *psaA* and *psaB*) were found to have elevated associated Ka/Ks ratios.

Stromal domains of peridinin plastid proteins have elevated Ka/Ks ratios

Individual Ka/Ks ratios were calculated for each residue within each plastid gene, for dinoflagellates, non-dinoflagellates, and the common ancestor of all studied dinoflagellates (Table S11). The distribution of residues with elevated Ka/Ks was biased towards specific regions of individual proteins. For example, within *psbB*, 59 codons have elevated associated Ka/Ks ratios either in the common ancestor of all studied dinoflagellates (positions at the centre of a block of ten residues with aggregate Ka/Ks >1, and Ka/Ks significantly greater than that calculated for the equivalent non-dinoflagellate residue; Z-test, $P< 0.05$), or within the dinoflagellates (same criteria, but with Ka/Ks > 0.5; Fig. S25). All of these codons are predicted to encode stromal- or luminal-facing *psbB* residues (Fig. S25).

Across the entire dataset, 171 of the 331 residues identified to have elevated Ka/Ks ratios in the common ancestor of all studied dinoflagellates were located on predicted stromal faces of plastid proteins (Fig. 5, panel B; Fig. S26, panel A; Tables S11, S12). This is significantly greater than the number (123) expected through a random distribution of residues (chi-squared, $P= 6.43 \times 10^{-10}$; Table S11). The elevated Ka/Ks ratios on stromal residues were also identified in calculations performed with alignments recoded to eliminate third position substitutions and codon usage bias (Fig. S26; Supplementary Results, Section 9).

The same trends were not directly observed within the dinoflagellates, where the number of stromal residues with elevated Ka/Ks was in fact slightly fewer than expected (202/582 residues; expected number 216; Fig. S26, panel B; Tables S11, S12). However, this may be influenced by the low Ka/Ks ratios observed in dinoflagellates for AtpA (0.251) and AtpB (0.254; Fig. 5, panel A; Table S11), which are the only two proteins encoded in peridinin dinoflagellate plastids to be entirely extrinsic to the thylakoid membrane, hence entirely stromal-facing (Walker 2013). Excluding AtpA and AtpB, the number of residues with elevated Ka/Ks in dinoflagellates that are predicted to face into the plastid stroma was extremely greater than expected (172/ 552 residues with elevated Ka/Ks; expected number of residues 29; P=0; Fig. S26, panel B; Tables S11, S12). Similar to the dinoflagellate ancestor,

the enrichment in dinoflagellate stromal Ka/Ks was confirmed through alignment recoding to be genuine, rather than a result of mutation rate saturation or codon usage bias (Supplementary Results, Section 9).

Discussion

We have identified and analysed previously unannotated sequences for plastid-encoded transcripts in peridinin dinoflagellates. These data constitute over three times the number of previously annotated peridinin dinoflagellate plastid sequences, and increases the number of complete peridinin plastid protein-coding datasets fivefold (Fig. 1). Particularly large numbers of plastid transcript sequences were identified from the MMETSP transcriptome datasets (Fig. 1). This may be due to the presence of the 3' poly(U) tail on dinoflagellate plastid transcripts, which has previously been speculated to enable the enrichment of plastid transcript sequences in poly(A) enriched RNA libraries (Wang and Morse 2006) such as those used for generation of the MMETSP libraries (Keeling, et al. 2014), and is corroborated by the presence of poly(U) tails on many of the transcripts we identified (Fig. S1; Table S1).

Almost all of the dinoflagellates investigated appear to possess the same twelve plastid-encoded protein-coding genes previously identified (Howe, et al. 2008; Barbrook, et al. 2014; Mungpakdee, et al. 2014). We found no evidence for relocation of any of these sequences to the nucleus in any species (Fig. S1; Table S1) or the retention of other plastid-derived sequences in any peridinin plastid (Fig. S2; Table S2). We additionally found only limited evidence for the presence of laterally acquired genes in the plastids of peridinin dinoflagellates, or for the conservation of other novel ORFs previously identified in individual peridinin plastid lineages across multiple species (Fig. S8; Table S2). While we cannot formally exclude that other plastid transcripts (transcripts that do not receive a 3' poly(U) tail hence are unlikely to be present in poly(A)-enriched libraries) are produced, our data indicates that the twelve previously identified protein-coding genes represents the ancestral protein-coding component of dinoflagellate plastids.

These data have allowed us to produce a well-resolved reference tree for the branching relationships between major clades of peridinin dinoflagellates (Fig. 2; Fig. S9; Fig. S10). Several novel phylogenetic relationships were uncovered. For example, *Pelagodinium beii* (previously *Gymnodinium beii*) grouped with the Prorocentrales with reasonable support (100/64%) whereas previous studies based on single-gene phylogenies placed it within the Suessiales (Siano, et al. 2010; Decelle, et al. 2014). Similarly, sister-group relationships between *Prorocentrum* and the Peridiniales, and the monophyletic clade of Gonyaulacales, Suessiales and *Heterocapsa* (Fig. 2), have not to our knowledge been previously described (Hoppenrath and Leander 2010; Bachvaroff, et al. 2014; Gavelis, et al. 2015). Each of these novel relationships were also recovered using modified alignments from which long branches, individual genes, gapped positions, or fast-evolving sites were removed (Fig. S9; Fig. S10). We accordingly conclude that the relationships obtained within our dataset are probably genuine, and not the artifact of fast sequence evolution or recent discrete changes (to plastid translation tables, which might bias the conceptual translations obtained, but presumably should be removed in the fast site analysis) in individual dinoflagellate plastids.

We have correlated divergent changes to peridinin plastid transcript sequences to the branching relationships from the multigene phylogeny, allowing us to infer when these changes occurred. For these analyses, we have focused exclusively on the sequences and conceptual translations of plastid transcripts, which provide an understanding of the aggregate consquences of divergent gene evolution and transcript editing on peridinin

plastids (Zauner, et al. 2004; Bachvaroff, et al. 2006; Dorrell and Howe 2015). As many of the sequences identified from MMETSP possess poly(U) tails (Fig. S1; Table 1), and the presence of the poly(U) tail is associated with the completion of transcript editing in dinoflagellate species (Dang and Green 2009; Dorrell, et al. 2016), we presume that the majority of the sequences identified in this study probably have been edited to completion, as opposed to representing unedited precursor transcripts.

First, we have identified changes to peridinin plastid sequence composition across the dinoflagellates. These include changes to plastid sequence GC content and codon usage (Fig. S11; Fig. S12), although many of these changes are either shared with relatives (such as the elevated third position GC content in the chromerid *Vitrella*; Fig. S11), or appear largely to consist of changes specific to individual dinoflagellate lineages (such as the changes to plastid codon usage frequencies, which are much more numerous in extant dinoflagellate species than in the inferred sequences of their last common ancestor; Fig. S12). We could not find convincing evidence for changes to the plastid translation table across the dinoflagellates (Fig. S13). While we cannot exclude alternative hypotheses (for example, independent increases in third position GC content in the lineages giving rise to *Vitrella*, and in a common ancestor of all dinoflagellates), we cannot find sufficient evidence for major changes to plastid sequence composition, or translation, occurring in the dinoflagellate common ancestor.

Next, we have identified widespread changes to the translation products of plastid sequences in dinoflagellates. These include large numbers of synonymous substitutions, and greatly elevated Ka/Ks ratios in the dinoflagellates (Fig. 3; Fig. S14; Fig. S16)). The elevated Ka/Ks ratios substitutions are unlikely to be explained by changes in codon preference (Fig. S11; Fig. S16), or saturation of synonymous substitution rates at third-position sites in dinoflagellate plastids (Fig. S17; Fig. S18; Fig. S19). It is possible that other factors related to sequence composition (e.g., transitory changes in codon preference in individual dinoflagellate lineages, or saturation of synonymous substitution rates at first- and second-position sites) may have contributed to the elevated Ka/Ks ratios observed; however, we suggest that the most parsimonious explanation for the remaining elevated Ka/Ks ratios observed in dinoflagellate plastids are changes in plastid selection pressure throughout their evolution, following their divergence from other plastid lineages, and prior to the radiation of extant species. Although previous studies have posited changes in selection pressure on certain peridinin plastid sequences (Shalchian-Tabrizi, et al. 2006), this is to our knowledge the first evidence that indicates that selective events have played a widespread role in dinoflagellate plastid evolution.

We have additionally identified changes that have occurred in individual dinoflagellate lineages since their radiation (Fig. 3; Fig. 4; Figs. S20-S23). We have found extremely elevated Ka/Ks ratios in pairwise comparisons between dinoflagellates (Fig. 3; Figs. S20; S21), and map multiple acquisitions of alternative translation initiation codons (Fig. S22), in-frame insertion and deletions (Fig. S23), and the loss of otherwise conserved *atpA* residues to the dinoflagellate tree (Fig. 4). These events have occurred progressively, with the majority of the dinoflagellate clades being marked by discrete changes to plastid sequence (Fig. 4), and contrasts with the much more conservative evolution observed in other plastid lineages (Tables S8, S9). The divergent evolutionary events observed in this study await detailed biochemical characterisation. For example, it will be interesting to determine experimentally whether peridinin dinoflagellates utilise the alternative translation initiation codons identified in this study, or whether there are further translation initiation codons with weaker similarity to ATG in peridinin plastids. This could be accomplished, for example,

by proteomic characterisation of the N-termini of peridinin plastid proteins (Huesgen, et al. 2013). Regardless, our data show that peridinin plastid sequences have not remained static, but have continued to diverge from one another to a remarkable extent.

We have found evidence that different peridinin plastid lineages have evolved in different manners since their radiation. For example, we observe elevated minimum pairwise substitution rates in some Peridinialean/ Prorocentralean species (e.g., *Brandtodinium, Pelagodinium*; Fig. 3), which is corroborated by the rather long branch lengths associated with these species in the multigene tree (Fig. 2; Fig. S9), and appears to explain at least partially explain the high Ka/Ks ratios observed for these species (Fig. S20). We also observe extremely high maximum pairwise Ka/Ks ratios within members of the Gonyaulacales and Suessiales, but in contrast these occur alongside relatively low pairwise substitution rates, suggesting that they have resulted from a change in plastid selective pressure in these lineages (Fig. 3; Fig. S21). Similarly, We have found that specific nodes on the peridinin plastid tree are marked by the origins of large numbers of alternative initiation codons (the common ancestor of *Scrippsiella hangoei* and *Peridinium aciculiferum*; Fig. 4), indels (the common ancestor of *Amphidinium* and *Heterocapsa*; Fig. 4), or changes to conserved *atpA* residues (the common ancestor of *Symbiodinium*; Fig. 4). It remains to be determined why the plastids of specific dinoflagellate lineages and not others are unusual, although we note that many of the most divergent species within our dataset have symbiotic life strategies (for example, *Pelagodinium* is an endobiont of foraminiferans (Siano, et al. 2010), whereas *Brandtodinium* is a radiolarian symbiont (Probert, et al. 2014)). More taxonomically detailed comparisons of plastid evolution in endosymbiotic dinoflagellates to their free-living relatives may provide insights into whether symbiosis has driven divergent plastid evolution within the dinoflagellates.

More globally, it remains to be resolved why the peridinin dinoflagellates have undergone such divergent plastid evolution compared to other lineages. Notably, we show that residues with elevated Ka/Ks ratios in peridinin plastids are concentrated, both in extant dinoflagellates and in the dinoflagellate common ancestor, on photosystem I subunits, and stromal regions of plastid proteins, which cannot be explained by changes to codon usage or third position substitution rate (Fig. 5; Fig. S24; Fig. S25; Fig. S26). While divergent evolution has previously been reported in the photosystem I sequences in individual peridinin plastid lineages (Bachvaroff, et al. 2006; Shalchian-Tabrizi, et al. 2006; Pochon, et al. 2014), this is to our knowledge the first evidence that this bias in evolutionary events is conserved throughout peridinin plastids, from the origin of dinoflagellates to extant species.

Similar, albeit less extreme divergent evolution events to those described in this study have been observed in nuclear genes encoding plastid-targeted proteins (Bachvaroff, et al. 2006; Mungpakdee, et al. 2014) and non-plastid proteins (Kim, et al. 2011) in peridinin dinoflagellates. It will be interesting to determine if the divergent evolution events observed in nucleus-encoded and plastid-encoded genes are linked; for example if the nucleus-encoded components of dinoflagellate photosystem I, or nucleus-encoded proteins likely to interact with the stromal faces of plastid-encoded dinoflagellate proteins, likewise possess specifically elevated Ka/Ks ratios. Conserved trends in the evolution of nucleus and plastid-encoded proteins in dinoflagellates might arise as a result of compensatory evolution events to maintain plastid physiology (for example, maintaining plastid redox state for plastid physiology and gene regulation (Allen 1993; Puthiyaveetil, et al. 2008), or balancing cyclic and linear electron flow to meet the physiological requirements of individual lineages (Reynolds, et al. 2008), or as a result of divergent selection on plastid proteins to accommodate some of the more unusual nucleus-encoded proteins and structures present

in peridinin plastids (for example, the dinoflagellate pyrenoid (Nassoury, et al. 2005; Siano, et al. 2010), or the peridinin pigment-binding protein complexes (Haxo, et al. 1976)). Ultimately, understanding the relationships between peridinin plastid sequences, evolution and physiology may provide valuable insights into the biology of this unusual and ecologically important lineage.

Materials and Methods

Assembly and annotation of previously known peridinin plastid sequences

Nucleotide sequences corresponding to the twelve protein-coding genes (*atpA, atpB, petB, petD, psaA, psaB, psbA, psbB, psbC, psbD, psbE, psbI*) that are found on the three effectively complete peridinin plastid genomes (*Amphidinium carterae* CCAP1102/6; *Symbiodinium sp*., Clade C3, and Clade Mf) (Nisbet, et al. 2004; Barbrook, et al. 2014; Mungpakdee, et al. 2014), were assembled from peridinin dinoflagellate EST libraries within NCBI and the Marine Microeukaryote Transcriptome Sequencing Project (MMETSP) (Keeling, et al. 2014). Sequences were identified by tBLASTn searches, using annotated plastid protein sequences from peridinin dinoflagellates and their close relatives (listed in Table S1). Where possible, the predicted translation products of plastid transcript sequences were used, instead of the translation products of plastid gDNA sequences, due to the presence of plastid transcript editing in some dinoflagellates (Zauner, et al. 2004; Dorrell and Howe 2015).

Sequences that matched the queries with expect values of below $1x10^{-5}$ were selected, and were searched using BLASTx against the entire NCBI database. Sequences that yielded a top hit against another peridinin dinoflagellate sequence were retained for subsequent analysis. Sequences that were excluded on the basis of being of probable non-dinoflagellate origin are listed in Table S1. In the case of *psbI*, for which peridinin dinoflagellate sequences are known to be highly divergent (Nisbet, et al. 2004), an initial expect value threshold <1.0 was used to identify possible dinoflagellate orthologues, and these orthologues were used in turn as query sequences for a second round of reciprocal tBLASTn/BLASTx searches against each dataset, to identify even the most divergent sequences present.

To investigate whether novel plastid ORFs, and genes proposed to have been acquired by lateral transfer from Bacteroidetes are conserved across peridinin plastids, similar BLAST searches were performed with translation products of the four predicted ORFs (*ORF1, ORF2, ORF3,* and *ORF4*) described in the *Amphidinium carterae* plastid genome (Barbrook, et al. 2001; Nisbet, et al. 2004; Barbrook, et al. 2006), and the proposed plastid *rpl28* and *rpl23* genes from *Pyrocystis lunula* and *ycf24* and *ycf16* sequences from *Ceratium horridum* (Moszczynski, et al. 2012). For these genes, sequences were retained if the top reciprocal BLAST hit was homologous to the genes in question, regardless of evolutionary affinity.

Sequences that passed the reciprocal BLASTx search were assembled into contigs using GeneIOUS v 4.76 (Kearse, et al. 2012). Each nucleotide sequence was searched manually for possible 5' spliced leader sequences (Zhang, et al. 2007), and uninterrupted 3' poly(T) and poly(A) tracts of more than 3 bp length, which might respectively correspond to transcript poly(U) and poly(A) tails (Wang and Morse 2006). The predicted translation products of each contig were inspected for the presence of possible plastid-targeting sequences using SignalP version 3.0 (Bendtsen, et al. 2004) and ASAFind (Gruber, et al. 2015), and for mitochondria-targeting sequences using TargetP version 1.1 (Emanuelsson, et al. 2007). Annotated copies of each novel peridinin plastid transcript (including information on the presence of possible

poly(U) tails, poly(A) tails, spliced-leader sequences, and plastid –targeting sequences) are included in Table S1.

Global identification and annotation of peridinin plastid transcripts

To determine whether further, previously undocumented ORFs are conserved across peridinin plastids, an independent, top-down search was performed for the entire dinoflagellate sequence library. We focused on transcripts possessing a 3' poly(U) tail, as this feature is believed to be uniquely associated with plastid transcripts in dinoflagellates (Dorrell and Howe 2012), and polyuridylylated transcripts have previously been indicated to be enriched by poly(A) selection, hence may be present in poly(A)-selected transcriptome datasets such as MMETSP (Wang and Morse 2006; Keeling, et al. 2014).

For this analysis, all sequences that terminated at either end in a potential poly(T) sequence (>3bp) from all RNA libraries were extracted, and filtered to remove sequences that either contained a more plausible reverse complement of a poly(A) sequence (length of poly(A) sequence ≥ poly(T) sequence), or a spliced-leader (≥ 6 bp spliced-leader sequence) on the other transcript end. Parallel libraries were constructed, containing all dinoflagellate sequences that by the same criteria possessed either a poly(A) tail, or spliced-leader sequences, and the remaining poly(T)-containing sequences were searched against these libraries using tBLASTx. Poly(T) sequences that were contiguous with other poly(A) or spliced-leader containing sequences (as judged by BLAST best hit; expect value < $1x10^{-5}$, > 90% similarities) were likewise removed from the dataset.

Next, to identify sequences encoding proteins of probable plastid function, the filtered poly(T)-containing sequences were oriented such that the poly(T) sequence was located on the transcript 3' end, and searched by BLASTx against a composite protein dataset, consisting of the complete protein sequences encoded in the nuclear and plastid genomes of the model diatom *Phaeodactylum tricornutum* (Oudot-Le Secq, et al. 2007; Bowler, et al. 2008), and all dinoflagellate plastid sequences previously identified by the BLAST searches detailed above. Sequences that yielded a top hit in a reverse orientation reading frame (i.e. such that the poly(T) sequence could not be located on a transcript 3' end), or were identified to correspond to an internal region of a transcript (judged if the 3' end of the transcript alignment corresponded to <90% the full length of the subject length), that yielded a top hit against a known dinoflagellate plastid protein, or that yielded a top hit against a *Phaeodactylum* nucleus-encoded protein (expect value < $1x10^{-5}$) were removed.

The longest ORFs from sequences that were found to possess plausible 3' ends, and were homologous to plastid-encoded proteins in *Phaeodactylum* that have not previously been identified in peridinin dinoflagellate plastids were extracted, and searched for the presence of plastid- and mitochondria-targeting sequences as above. The sequences for which no plausible BLAST hit was found (and thus may contain novel plastid ORFs), were searched against one another using reciprocal tBLASTx/tBLASTx searches. Sequences that were found to match one another with a reciprocal BLAST hit expect value of < $1x10^{-5}$ (and thus might correspond to conserved novel plastid proteins) were retained, assembled into contigs, and aligned using GeneIOUS v 4.76, as above. The final annotations for each poly(T)-containing sequence, as well as full nucleotide and protein sequences for the possible novel plastid ORFs identified, are presented in Table S2.

Phylogenetic analysis

To assess the phylogenetic relationships between the different dinoflagellates studied, conceptual translations were generated for each novel peridinin plastid sequence identified from the MMETSP libraries (Keeling, et al. 2014), using a standard translation table. This and all subsequent analyses conducted throughout the remainder of the study were restricted to proteins generated through the conceptual translations of plastid transcript sequences, to minimise complications resulting from the comparison of (edited) transcript sequences and (unedited) gene sequences, and to reflect most accurately the probable eventual translation products of each plastid sequence.

The conceptual dinoflagellate protein sequences for each gene were aligned, using twenty iterations of MUSCLE v 3.8 (Edgar 2004), to a reference set of orthologous sequences from fifteen other plastid lineages, including a chromerid (*Vitrella brassicaformis*) that represents the closest documented plastid-containing relative of the peridinin dinoflagellates (Janouskovec, et al. 2010). Orthologues from other alveolate plastids (those of apicomplexans, and the chromerid *Chromera velia*) were not included, due to the loss of photosynthesis genes (in the case of apicomplexans), or the exceptionally divergent nature of the plastid genome, including changes in the plastid translation table (in the case of *C. velia*) (Moore, et al. 2008; Janouskovec, et al. 2013). The *petD* sequence from *Ostreococcus tauri* was excluded from this and all subsequent analyses, as it is not present within the plastid genome (Derelle, et al. 2006; Robbens, et al. 2007).

The alignments were manually corrected, trimmed to remove all positions at which the most common non-ambiguous identity was a gapped position, and concatenated. For single-gene trees of *rpl28, rpl33, ycf16* and *ycf24* sequences, similar alignments were generated, in this case containing all of the novel sequences identified, the previously annotated *Pyrocystis lunula* and *Ceratium horridum* sequences (Moszczynski, et al. 2012), and a representative sample of sequences from different plastid and bacterial lineages, and trimmed as before. All of the multiple- and single-gene alignments generated for inference of phylogenetic trees are provided in Table S3.

Bayesian inference of single gene alignments was carried out using PhyloBayes v3.3 (Lartillot, et al. 2009), with default settings and the LG + Γ model. Two chains were run in parallel, using the automatic stopping rule such that sampling was conducted every 100 points until the maximum difference was ≤ 0.1 and the effective size ≥ 100. Bayesian analyses of concatenated alignments were performed using MrBayes 3.2.6 hosted on the CIPRES Science Gateway webportal (Ronquist, et al. 2012; Miller, et al. 2015). For each dataset, two independent runs were performed, each comprising four chains for 1,000,000 MCMC generations, sampling every 1,000 points and selecting the first quarter as burn-in for the consensus. The burn-in was selected such that the resulting standard deviation of split frequencies was < 0.06 for all datasets, and that the log likelihood of the cold chain had reached a stable plateau. Analyses were run with model mixing and Γ distributed rate categories with an additional invariant category.

All Maximum Likelihood analyses were carried out using RAxML v8.1.17 (Stamatakis 2014), under the LG + Γ model, as hosted on CIPRES. For analyses of both Bayesian and Maximum Likelihood inference, the –b option was used to conduct 500 non-parametric bootstraps. These were then subsequently assembled into a consensus tree using the consense program of the PHYLIP package (v3.695) on default settings. Bipartitions present in both trees were mapped onto the Bayesian topology. Trees based solely on Maximum Likelihood were run using the –f a option, such that bootstraps were automatically mapped onto the Maximum Likelihood estimate of the tree from the same run. In all cases, gamma distributed rates

were modeled under a discrete model with four rate categories. Newick format tree outputs for each alignment, under each condition tested, are provided in Table S4.

Alignments from which individual genes or all gapped positions were removed were constructed using GeneIOUS. Ten further alignments were constructed from which the species corresponding to the ten longest branches obtained in the initial Bayesian consensus tree analysis were serially removed. Finally, the evolutionary rate associated with each position within the alignment was calculated using TIGER, with the -b 100 option (Cummins and McInerney 2011). Fifteen alignments were constructed, from which the twenty-five fastest evolving categories were serially removed. All modified alignments were tested using RAxML using the same conditions as previously defined. Modified alignments and TIGER output data for fast site identification are supplied in Table S3.

Codon usage comparisons

Changes in nucleotide sequence composition across the peridinin dinoflagellates were assessed over the six genes (*psaA, psaB, psbA, psbB, psbC,* and *psbD*) for which sequences had been identified in each dinoflagellate MMETSP library studied. Due to the probable sequence contamination present, the *Symbiodinium* Clade A library was excluded from this and all subsequent analyses (Fig. 2). First, nucleotide alignments of each gene were generated, and manually trimmed to remove insertions specific to dinoflagellate lineages; these alignments were concatenated to produce an alignment of all six genes (Table S3). GC content and codon usage frequencies of each sequence within the alignment were calculated using PAML v4.8 (Yang 2007), using a standard genetic code. Amino acid composition of each species was quantified by summing the different codon frequencies obtained. To assess the relative codon usage amino acid composition biases associated with each species, the average frequency of each codon, and each amino acid, were calculated for all species in the alignment. The sum of squares of difference between the frequencies of each codon, or amino acid for each species, and the alignment-wide mean values, were calculated, and then ranked in ascending order. All statistics pertaining to codon and amino acid frequency are provided in Table S5.

To assess the distribution and form of translationally invariant sites in dinoflagellate plastids, untrimmed nucleotide alignments were generated for each gene. dN and dS ratios were calculated for each site, within each gene, for dinoflagellate sequences only, using PAML v4.8 under a standard genetic code. All codon positions within each alignment with an observed Ka of 0 (i.e. no non-synonymous substitutions) within the dinoflagellates were extracted, and are shown in Table S6.

Finally, to assess which of the trends identified might represent ancestral changes to dinoflagellate plastid translation, predicted ancestral sequences of each plastid sequence were generated by regression using PAML as before (Table S3). Two sequences were generated, one corresponding to the common ancestor of all studied dinoflagellates included in this study, and one corresponding to the common ancestor of dinoflagellates and *Vitrella brassicaformis*, the closest sister-group to dinoflagellates within the alignment. Codon frequencies associated with each ancestral sequence were calculated using PAML v4.8 as before, and compared to one another.

Nucleotide sequence substitutions

Pairwise Ka/Ks ratios and total number of pairwise substitutions were calculated using the

gap-free concatenated nucleotide alignment, for each possible combination of species, with KaKs Calculator version 2.0, using the Model Averaging method (which performs weighted calculations using eight different substitution matrices, based on the relative likelihood of each matrix to explain the observed sequences) and a standard genetic code (Wang, et al. 2010). Pairwise Ka/Ks calculations and the total number of substitutions for each species pair are provided in Table S7.

To investigate the different factors underpinning substitution rates within the alignment, total numbers of pairwise substitutions and Ka/Ks ratios were correlated against (1) the third-position GC content, (2) the codon usage bias (defined as above) and (3) the amino acid composition bias for each species pair. Correlation calculations were performed for both the average and modular differences in each variable for each species pair, and were performed for both the absolute substitution rates and Ka/Ks ratios observed, and the rank of each value within the entire dataset. Third-position GC, amino acid composition bias and codon usage bias values for each species are provided in Table S5, and values for each species pair are provided in Table S7.

Mutation rate saturation was assessed using the DAMBE software package (Xia, et al. 2003). The concatenated gap-free alignment was separated into first-, second-, and third-codon positions only, and separate calculations were performed for each codon position. A separate series of calculations were performed for each codon position using alignments consisting only of the dinoflagellate sequences from the first-, second- or third-position alignments. For each alignment and codon position, the proportion of invariant sites was first calculated, using the tree topology obtained using the multigene phylogeny (Table S4), under otherwise default conditions. Substitution rates were then assessed at each codon position through each alignment using Xia's test, with the empirically verified invariant site frequencies (Xia, et al. 2003; Xia 2013), and otherwise under the default conditions.

To determine whether any of the above factors have had significant effects on the total numbers of pairwise substitutions observed, the total number of pairwise substitutions and Ka/Ks ratios were calculated for alignments that were manually recoded in several ways (all modified alignments provided in Table S3). To investigate the affect of GC content shifting on substitution rates, total number of pairwise substitutions were calculated for alignments that had been RY-recoded (Ishikawa, et al. 2012). Two RY-recoded alignments were produced: one alignment in which all adenosines were replaced by guanosines and all thymidines were replaced by cytosines, and one for which the converse substitutions were performed. To investigate the different roles of each codon positions on substitution rates, total number of pairwise substitutions were calculated for the separate first, second and third position alignments previously generated for inspection with DAMBE, and two further alignments consisting of third codon positions only that had been manually RY-recoded as above (Table S3). To investigate the significance of third position mutations on Ka/Ks, ratios were calculated for two alignments that had been manually recoded so that all third codon positions were replaced either by adenosines, or by guanosines (Table S3). These two substitutions were performed as they preserve the possibility for synonymous first-position substitutions associated with leucine (YTR) and arginine codons (RGR), whereas manual recoding of the third position to pyrimidine nucleotides would eliminate all synonymous single substitutions associated with six-fold degenerate codons under a standard genetic code (Crick 1968; Inagaki, et al. 2004). Finally, to investigate the effects of changes in dinoflagellate plastid codon preference on Ka/Ks and substitution rates, calculations were performed for an alignment which was globally recoded so that 19 codons, which occur at significantly lower frequencies in dinoflagellates than in non-dinoflagellates (one-way

ANOVA, P<0.05), were universally replaced with 19 synonymous codons that occur at elevated frequencies in dinoflagellates (Table S5). Pairwise substitution value calculations for all manually recoded alignments are provided in Table S7.

Translation initiation site identification

Possible alternative translation initiation codons in peridinin plastids were identified from nucleotide sequence alignments of each sequence that contained all of the dinoflagellate sequences studied, and orthologues from each of the non-dinoflagellate reference sequences used for construction of the multigene phylogeny. The 5' end of each nucleotide sequence was trimmed so that it started from the first in-frame termination codon within the 5' UTR, or (if no such codon existed) so that the coding sequence was located within reading frame position +1.

To determine which alternative translation initiation codons are utilised in peridinin plastids, the most probable initiation codon was identified in the N-terminal coding region of each dinoflagellate sequence using a custom-built automated pipeline (Text S1). As the presence of Shine-Dalgarno sequences on peridinin plastid genes remains unclear (Zhang, et al. 1999; Dang and Green 2009), initiation codons were searched for in the entire possible 5' coding region of each sequence. This was defined as the region extending upstream from the first residue that was identical in 75% of the non-dinoflagellate reference sequences. The most probable initiation codon was taken to be ATG, if such a codon were present. If not, the sequence was screened for codons that shared two bases in common with ATG, which has been shown experimentally in other plastid lineages to be sufficient to permit translation initiation (Chen, et al. 1995), alongside GTA, which has been documented to function as an alternative initiation codon in Proteobacteria (Kim, et al. 2008), and has previously been proposed to be used as an alternative translation initiation codon in *Amphidinium* (Barbrook and Howe 2000; Barbrook, et al. 2001). The nearest codon to the consensus initiation site in the non-dinoflagellate reference species (hence would produce the ORF with the greatest homology to the expected protein sequence) was then taken as the most probable translation initiation site. These data, including full lists of all codons deemed to be possible initiation codons for each gene and in each species, as well as the most probable initiation codons, are provided in Table S8.

Indel analysis

Indels within peridinin plastid sequences were identified using untrimmed protein sequence alignments generated for each gene, as above. Pairwise comparisons were performed between each dinoflagellate sequence and non-dinoflagellate sequence previously used for the construction of the multigene alignment, using a custom built Python script that automatically detects and reports both insertions and deletions (Text S2). Sequences were counted as insertions only if they occurred in at least one dinoflagellate but were absent from all of the non-dinoflagellate reference sequences, or as deletions if they were absent from at least one dinoflagellate, but present in every reference sequence examined.

Each predicted indel was confirmed by visual inspection of the alignment. Positions that were within 5 amino acids of sequence N- or C-termini, and patterns of alternating insertions and deletions that indicated poor alignment of the sequence were rejected. Tabulated indels identified for each species, and the inferred phylogenetic origin point (as defined using the multigene tree topology) of each indel is shown in Table S9.

Changes to conserved atpA residues

Changes to residues that were conserved in all other plastid lineages were identified by visual inspection of the untrimmed atpA protein sequence alignment containing all of the dinoflagellate sequences studied, and orthologues from each of the non-dinoflagellate reference sequences used for generation of the multigene phylogeny. Residues were deemed to be conserved in non-dinoflagellate lineages only if they were found in all, or all but one, of the species studied, including the closest relative to the dinoflagellates within the alignment (*Vitrella brassicaformis*). The full list of substitutions to conserved AtpA residues for each species, alongside the inferred origin points for each substitution, is provided in Table S10.

Site-specific substitution rates

Substitution rates for each residue within each sequence were calculated using the previously constructed single-gene nucleotide alignments. Each sequence was trimmed to remove indels that were specific to dinoflagellates, and Ka/Ks ratios were calculated for each site using PAML v 4.8, site model constraint 2, and a standard genetic code (Yang 2007). The consensus multigene tree generated above was used as the reference tree topology. Separate alignments were constructed for each sequence in which the codon third positions were either manually recoded to adenosines or to guanosines, or in which all codon positions were manually altered to reflect the predominant codon usage patterns observed in dinoflagellates, as detailed above. Raw and modified alignments for each sequence are provided in Table S3, and the Newick format trees used for each analysis in Table S4.

To allow comparisons between the evolutionary rates observed for dinoflagellate and non-dinoflagellate species, separate values were calculated for each site, using only dinoflagellates, and using only non-dinoflagellate sequences. To identify substitution events that occurred at the divergence of peridinin dinoflagellates from other plastids, values were also calculated for each gene for the previously inferred sequence for the dinoflagellate common ancestor, when compared directly to the sequence for the common ancestor of dinoflagellates and *Vitrella* (Table S3). Individual rate calculations for each gene, including gene length, are provided in Table S11.

Averaged Ka/Ks ratios were then calculated over a sliding window of eleven residues in each sequence. For regions of sequence which through this approach contained no synonymous substitutions, the sliding window was expanded to include the nearest codon upstream, and the nearest codon downstream, to have undergone synonymous substitutions. The residues with Ka/Ks >1 in the dinoflagellate common ancestor, or Ka/Ks > 0.5 within dinoflagellates were inspected for localisation (transmembrane, stromal or luminal-facing), and function (interaction with other plastid-encoded subunits, or with nucleus-encoded subunits and cofactors), as inferred by alignment to annotated PDB sequences from the plastid of *Arabidopsis thaliana* (Sato, et al. 1999), and the model cyanobacterium *Thermosynechococcus elongatus* (Kamiya and Shen 2003). Separate Ka/Ks calculations were also performed and annotated for the previously generated codon re-optimised, third-position adenosine- and third-position guanosine recoded alignments (Table S3), with substitution values for each alignment provided in Tables S12-S15.

Data Deposition

All sequences and supplementary tables referenced in this study are publically accessible from the University of Cambridge dserve server (https://www.repository.cam.ac.uk/handle/1810/252774).

Acknowledgments

This work was supported by a "Research in Paris" post-doctoral fellowship [to RGD] from the Mairie de Paris (grant number 2014/89). Work in the Bowler lab is supported by the ERC Advanced Award Diatomite.

References

Allen JF. 1993. Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes. J Theor Biol 165:609-631.
Bachvaroff TR, Concepcion GT, Rogers CR, Herman EM, Delwiche CF. 2004. Dinoflagellate expressed sequence tag data indicate massive transfer to the nuclear genome sequence. Protist 155:65-78.
Bachvaroff TR, Gornik SG, Concepcion GT, Waller RF, Mendez GS, Lippmeier JC, Delwiche CF. 2014. Dinoflagellate phylogeny revisited: Using ribosomal proteins to resolve deep branching dinoflagellate clades. Mol Phyl Evol 70:314-322.
Bachvaroff TR, Sanchez-Puerta MV, Delwiche CF. 2006. Rate variation as a function of gene origin in plastid-derived genes of peridinin-containing dinoflagellates. J Mol Evol 62:42-52.
Barbrook AC, Dorrell RG, Burrows J, Plenderleith LJ, Nisbet RER, Howe CJ. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae*. Plant Mol Biol 79:347-357.
Barbrook AC, Howe CJ. 2000. Minicircular plastid DNA in the dinoflagellate Amphidinium operculatum. Molecular and General Genetics 263:152-158.
Barbrook AC, Santucci N, Plenderleith LJ, Hiller RG, Howe CJ. 2006. Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. BMC Genom 7: 297.
Barbrook AC, Symington H, Nisbet RER, Larkum A, Howe CJ. 2001. Organisation and expression of the plastid genome of the dinoflagellate *Amphidinium operculatum*. Mol Genet Genom 266:632-638.
Barbrook AC, Voolstra CR, Howe CJ. 2014. The chloroplast genome of a *Symbiodinium* sp clade C3 isolate. Protist 165:1-13.
Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783-795.
Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456:239-244.
Butterfield ER, Howe CJ, Nisbet RER. 2013. An analysis of dinoflagellate metabolism using EST data. Protist 164:218-236.
Chen X, Kindle KL, Stern DB. 1995. The initiation codon determines the efficiency but not the site of translation initiation in *Chlamydomonas* chloroplasts. Plant Cell 7:1295-1305.
Crick FH. 1968. The origin of the genetic code. J Mol Biol 38:367-379.
Cummins CA, McInerney JO. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst Biol 60:833-844.
Dang Y, Green BR. 2009. Substitutional editing of *Heterocapsa triquetra* chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA. Gene 442:73-80.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, et al. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605.
Decelle J, Siano R, Probert I, Poirier C, Not F. 2014. Multiple microalgal partners in symbiosis with the acantharian Acanthochiasma sp. (Radiolaria). Symbiosis 58:233-244.
Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, et al. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci USA 103:11647-11652.
Dorrell RG, Drew J, Nisbet RE, Howe CJ. 2014. Evolution of chloroplast transcript processing in *Plasmodium* and its chromerid algal relatives. PLoS Genet 10:1004008.
Dorrell RG, Hinksman GA, Howe CJ. 2016. Diversity of transcripts and transcript processing forms in plastids of the dinoflagellate alga *Karenia mikimotoi*. Plant Mol Biol 90:233-247.
Dorrell RG, Howe CJ. 2012. Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. Proc Natl Acad Sci USA 109: 18879-18884.
Dorrell RG, Howe CJ. 2015. Integration of plastids with their hosts: lessons learnt from dinoflagellates Proc Natl Acad Sci USA 112: 10247–10254.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 32:1792-1797.
Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protocol 2:953-971.
Gavelis GS, White RA, Suttle CA, Keeling PJ, Leander BS. 2015. Single-cell transcriptomics using spliced leader PCR: evidence for multiple losses of photosynthesis in polykrikoid dinoflagellates. BMC Genom 16:528.
Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF. 2012. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. Curr Biol 22:2303-2312.
Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T. 2015. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. Plant J 81:519-528.
Hallegraeff GM. 2010. Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge. J Phycol 46:220-235.
Haxo FT, Kycia JH, Somers GF, Bennett A, Siegelman HW. 1976. Peridinin chlorophyll a proteins of dinoflagellate *Amphidinium carterae* (Plymouth 450). Plant Physiol 57:297-303.
Hiller RG. 2001. 'Empty' minicircles and petB/atpA and psbD/psbE (cytb(559) alpha) genes in tandem in *Amphidinium carterae* plastid DNA. FEBS Lett 505:449-452.
Hinder SL, Hays GC, Edwards M, Roberts EC, Waine AW, Gravenor MB. 2012. Changes in marine dinoflagellate and diatom abundance under climate change. Nat Climate Change. 2: 271-275.
Hoppenrath M, Leander BS. 2010. Dinoflagellate phylogeny as inferred from heat shock protein 90 and ribosomal gene sequences. PLoS One 5: 13220.
Howe CJ, Nisbet RER, Barbrook AC. 2008. The remarkable chloroplast genome of dinoflagellates. J Exp Bot 59:1035-1045.
Huesgen PF, Alami M, Lange PF, Foster LJ, Schröder WP, Overall CM, Green BR. 2013. Proteomic amino-termini profiling reveals targeting information for protein import into complex plastids. PLoS One 8:74483.
Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18:486.
Inagaki Y, Simpson AGB, Dacks JB, Roger AJ. 2004. Phylogenetic artifacts can be caused by leucine, serine, and arginine codon usage heterogeneity: dinoflagellate plastid origins as a case study. Systemat Biology 53:582-593.

Ishikawa SA, Inagaki Y, Hashimoto T. 2012. RY-Coding and non-homogeneous models can ameliorate the maximum-likelihood inferences from nucleotide sequence data with parallel compositional heterogeneity. Evol Bioinf 8:357-371.
Janouskovec J, Horák A, Oborník M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci USA 107:10949-10954.
Janouskovec J, Sobotka R, Lai DH, Flegontov P, Koník P, Komenda J, Ali S, Prásil O, Pain A, Oborník M, et al. 2013. Split photosystem protein, linear-mapping topology, and growth of structural complexity in the plastid genome of *Chromera velia*. Mol Biol Evol 30:2447-2462.
Janouškovec J, Tikhonenkov D, Mikhailov K, Simdyanov T, Aleoshin V, Mylnikov A, Keeling P. 2013. Colponemids represent multiple ancient alveolate lineages. Curr Biol, 23: 2546-2552.
Kamiya N, Shen JR. 2003. Crystal structure of oxygen-evolving photosystem II from *Thermosynechococcus vulcanus* at 3.7-A resolution. Proc Natl Acad Sci USA 100:98-103.
Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinf 28:1647-1649.
Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. Phil Trans R Soc Biol Sci 365:729-748.
Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol 12:1001889.
Kim S, Bachvaroff TR, Handy SM, Delwiche CF. 2011. Dynamics of actin evolution in dinoflagellates. Mol Biol Evol 28:1469-1480.
Kim SW, Jung WH, Ryu JM, Kim JB, Jang HW, Jo YB, Jung JK, Kim JH. 2008. Identification of an alternative translation initiation site for the *Pantoea ananatis* lycopene cyclase (*crtY*) gene in *E. coli* and its evolutionary conservation. Protein Expr Purif 58:23-31.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinf 25:2286-2288.
Leterme SC, Seuront L, Edwards M. Differential contribution of diatoms and dinoflagellates to phytoplankton biomass in the NE Atlantic Ocean and the North Sea. Mar Ecol Prog Ser 312:9.
Lin S. 2011. Genomic understanding of dinoflagellates. Res Microbiol 162:551-569.
Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, Passarotti M, Kaufman S, O'Leary MA. 2015. A RESTful API for access to phylogenetic tools via the CIPRES science gateway. Evol Bioinform Online 11:43-48.
Moore RB, Obornik M, Janouskovec J, Chrudimsky T, Vancova M, Green DH, Wright SW, Davies NW, Bolch CJS, Heimann K, et al. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. Nature 451:959-963.
Morse D, Salois P, Markovic P, Hastings JW. 1995. A nuclear-encoded form II RuBisCO in dinoflagellates. Science 268:1622-1624.
Moszczynski K, Mackiewicz P, Bodyl A. 2012. Evidence for horizontal gene transfer from Bacteroidetes bacteria to dinoflagellate minicircles. Mol Biol Evol 29:887-892.
Mungpakdee S, Shinzato C, Takeuchi T, Kawashima T, Koyanagi R, Hisata K, Tanaka M, Goto H, Fujie M, Lin S, et al. 2014. Massive gene transfer and extensive RNA editing of a symbiotic dinoflagellate plastid genome. Genome Biol Evol 6:1408-1422.
Nakamura Y. 1998. Growth and grazing of a large heterotrophic dinoflagellate, Noctiluca scintillans, in laboratory cultures. J Plankton Res 20:1711-1720.
Nassoury N, Morse D. 2005. Protein targeting to the chloroplasts of photosynthetic eukaryotes: getting there is half the fun. Biochimic Biophys Acta 1743:5-19.

Nassoury N, Wang Y, Morse D. 2005. Brefeldin a inhibits circadian remodeling of chloroplast structure in the dinoflagellate gonyaulax. Traffic 6:548-561.
Nelson MJ, Dang YK, Filek E, Zhang ZD, Yu VWC, Ishida K, Green BR. 2007. Identification and transcription of transfer RNA genes in dinoflagellate plastid minicircles. Gene 392:291-298.
Nelson MJ, Green BR. 2005. Double hairpin elements and tandem repeats in the non-coding region of Adenoides eludens chloroplast gene minicircles. Gene 358:102-110.
Nisbet RER, Koumandou VL, Barbrook AC, Howe CJ. 2004. Novel plastid gene minicircles in the dinoflagellate *Amphidinium operculatum*. Gene 331:141-147.
Oudot-Le Secq MP, Grimwood J, Shapiro H, Armbrust EV, Bowler C, Green BR. 2007. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. Molecular Genet Genom 277:427-439.
Pochon X, Putnam HM, Gates RD. 2014. Multi-gene analysis of *Symbiodinium* dinoflagellates: a perspective on rarity, symbiosis, and evolution. PeerJ 2:394.
Probert I, Siano R, Poirier C, Decelle J, Biard T, Tuji A, Suzuki N, Not F. 2014. *Brandtodinium* gen. nov. and *B. nutricula* comb. nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. J Phycol 50:388-399.
Puthiyaveetil S, Kavanagh TA, Cain P, Sullivan JA, Newell CA, Gray JC, Robinson C, van der Giezen M, Rogers MB, Allen JF. 2008. The ancestral symbiont sensor kinase CSK links photosynthesis with gene expression in chloroplasts. Proc Natl Acad Sci USA 105:10061-10066.
Reynolds JM, Bruns BU, Fitt WK, Schmidt GW. 2008. Enhanced photoprotection pathways in symbiotic dinoflagellates of shallow-water corals and other cnidarians. Proc Natl Acad Sci USA 105:13674-13678.
Robbens S, Derelle E, Ferraz C, Wuyts J, Moreau H, Van de Peer Y. 2007. The complete chloroplast and mitochondrial DNA sequence of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. Mol Biol Evol 24:956-968.
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539-542.
Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res 6:283-290.
Shalchian-Tabrizi K, Skanseng M, Ronquist F, Klaveness D, Bachvaroff TR, Delwiche CF, Botnen A, Tengs T, Jakobsen KS. 2006. Heterotachy processes in rhodophyte-derived secondhand plastid genes: implications for addressing the origin and evolution of dinoflagellate plastids. Mol Biol Evol 23:1504-1515.
Siano R, Montresor M, Probert I, Not F, de Vargas C. 2010. *Pelagodinium* gen. nov. and P. béii comb. nov., a dinoflagellate symbiont of planktonic foraminifera. Protist 161:385-399.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinf 30:1312-1313.
Walker JE. 2013. The ATP synthase: the understood, the uncertain and the unknown. Biochemical Society Transactions 41:1-16.
Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom Proteom Bioinf 8:77-80.
Wang YL, Morse D. 2006. Rampant polyuridylylation of plastid gene transcripts in the dinoflagellate *Lingulodinium*. Nucl Acids Rese 34:613-619.
Wisecaver JH, Hackett JD. 2011. Dinoflagellate genome evolution. Ann Rev Microbiol 65:369-387.
Xia X. 2013. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30:1720-1728.

Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Mol Phylogenet Evol 26:1-7.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496-503.
Zauner S, Greilinger D, Laatsch T, Kowallik KV, Maier UG. 2004. Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellate Ceratium horridum. FEBS Lett 577:535-538.
Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, Lin S. 2007. Spliced leader RNA trans-splicing in dinoflagellates. Proc Natl Acad Sci USA 104:4618-4623.
Zhang Z, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. Nature 400:155-159.
Ševčíková T, Horák A, Klimeš V, Zbránková V, Demir-Hilton E, Sudek S, Jenkins J, Schmutz J, Přibyl P, Fousek J, et al. 2015. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? Sci Rep 5:10134.

Allen JF. 1993. Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes. J Theor Biol 165:609-631.
Bachvaroff TR, Concepcion GT, Rogers CR, Herman EM, Delwiche CF. 2004. Dinoflagellate expressed sequence tag data indicate massive transfer to the nuclear genome sequence. Protist 155:65-78.
Bachvaroff TR, Gornik SG, Concepcion GT, Waller RF, Mendez GS, Lippmeier JC, Delwiche CF. 2014. Dinoflagellate phylogeny revisited: Using ribosomal proteins to resolve deep branching dinoflagellate clades. Mol Phyl Evol 70:314-322.
Bachvaroff TR, Sanchez-Puerta MV, Delwiche CF. 2006. Rate variation as a function of gene origin in plastid-derived genes of peridinin-containing dinoflagellates. J Mol Evol 62:42-52.
Barbrook AC, Dorrell RG, Burrows J, Plenderleith LJ, Nisbet RER, Howe CJ. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae*. Plant Mol Biol 79:347-357.
Barbrook AC, Howe CJ. 2000. Minicircular plastid DNA in the dinoflagellate Amphidinium operculatum. Molecular and General Genetics 263:152-158.
Barbrook AC, Santucci N, Plenderleith LJ, Hiller RG, Howe CJ. 2006. Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. BMC Genom 7: 297.
Barbrook AC, Symington H, Nisbet RER, Larkum A, Howe CJ. 2001. Organisation and expression of the plastid genome of the dinoflagellate *Amphidinium operculatum*. Mol Genet Genom 266:632-638.
Barbrook AC, Voolstra CR, Howe CJ. 2014. The chloroplast genome of a *Symbiodinium* sp clade C3 isolate. Protist 165:1-13.
Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783-795.
Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456:239-244.
Butterfield ER, Howe CJ, Nisbet RER. 2013. An analysis of dinoflagellate metabolism using EST data. Protist 164:218-236.
Chen X, Kindle KL, Stern DB. 1995. The initiation codon determines the efficiency but not the site of translation initiation in *Chlamydomonas* chloroplasts. Plant Cell 7:1295-1305.
Crick FH. 1968. The origin of the genetic code. J Mol Biol 38:367-379.
Cummins CA, McInerney JO. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst Biol 60:833-844.
Dang Y, Green BR. 2009. Substitutional editing of *Heterocapsa triquetra* chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA. Gene 442:73-80.
de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, et al. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605.
Decelle J, Siano R, Probert I, Poirier C, Not F. 2014. Multiple microalgal partners in symbiosis with the acantharian Acanthochiasma sp. (Radiolaria). Symbiosis 58:233-244.
Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, et al. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci USA 103:11647-11652.
Dorrell RG, Drew J, Nisbet RE, Howe CJ. 2014. Evolution of chloroplast transcript processing in *Plasmodium* and its chromerid algal relatives. PLoS Genet 10:1004008.
Dorrell RG, Hinksman GA, Howe CJ. 2016. Diversity of transcripts and transcript processing forms in plastids of the dinoflagellate alga *Karenia mikimotoi*. Plant Mol Biol 90:233-247.

Dorrell RG, Howe CJ. 2012. Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. Proc Natl Acad Sci USA 109: 18879-18884.
Dorrell RG, Howe CJ. 2015. Integration of plastids with their hosts: lessons learnt from dinoflagellates Proc Natl Acad Sci USA 112: 10247–10254.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 32:1792-1797.
Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protocol 2:953-971.
Gavelis GS, White RA, Suttle CA, Keeling PJ, Leander BS. 2015. Single-cell transcriptomics using spliced leader PCR: evidence for multiple losses of photosynthesis in polykrikoid dinoflagellates. BMC Genom 16:528.
Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF. 2012. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. Curr Biol 22:2303-2312.
Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T. 2015. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. Plant J 81:519-528.
Hallegraeff GM. 2010. Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge. J Phycol 46:220-235.
Haxo FT, Kycia JH, Somers GF, Bennett A, Siegelman HW. 1976. Peridinin chlorophyll a proteins of dinoflagellate *Amphidinium carterae* (Plymouth 450). Plant Physiol 57:297-303.
Hiller RG. 2001. 'Empty' minicircles and petB/atpA and psbD/psbE (cytb(559) alpha) genes in tandem in *Amphidinium carterae* plastid DNA. FEBS Lett 505:449-452.
Hinder SL, Hays GC, Edwards M, Roberts EC, Waine AW, Gravenor MB. 2012. Changes in marine dinoflagellate and diatom abundance under climate change. Nat Climate Change. 2: 271-275.
Hoppenrath M, Leander BS. 2010. Dinoflagellate phylogeny as inferred from heat shock protein 90 and ribosomal gene sequences. PLoS One 5: 13220.
Howe CJ, Nisbet RER, Barbrook AC. 2008. The remarkable chloroplast genome of dinoflagellates. J Exp Bot 59:1035-1045.
Huesgen PF, Alami M, Lange PF, Foster LJ, Schröder WP, Overall CM, Green BR. 2013. Proteomic amino-termini profiling reveals targeting information for protein import into complex plastids. PLoS One 8:74483.
Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18:486.
Inagaki Y, Simpson AGB, Dacks JB, Roger AJ. 2004. Phylogenetic artifacts can be caused by leucine, serine, and arginine codon usage heterogeneity: dinoflagellate plastid origins as a case study. Systemat Biology 53:582-593.
Ishikawa SA, Inagaki Y, Hashimoto T. 2012. RY-Coding and non-homogeneous models can ameliorate the maximum-likelihood inferences from nucleotide sequence data with parallel compositional heterogeneity. Evol Bioinf 8:357-371.
Janouskovec J, Horák A, Oborník M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci USA 107:10949-10954.
Janouskovec J, Sobotka R, Lai DH, Flegontov P, Koník P, Komenda J, Ali S, Prásil O, Pain A, Oborník M, et al. 2013. Split photosystem protein, linear-mapping topology, and growth of structural complexity in the plastid genome of *Chromera velia*. Mol Biol Evol 30:2447-2462.
Janouškovec J, Tikhonenkov D, Mikhailov K, Simdyanov T, Aleoshin V, Mylnikov A, Keeling P. 2013. Colponemids represent multiple ancient alveolate lineages. Curr Biol, 23: 2546-2552.
Kamiya N, Shen JR. 2003. Crystal structure of oxygen-evolving photosystem II from *Thermosynechococcus vulcanus* at 3.7-A resolution. Proc Natl Acad Sci USA 100:98-103.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinf 28:1647-1649.
Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. Phil Trans R Soc Biol Sci 365:729-748.
Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol 12:1001889.
Kim S, Bachvaroff TR, Handy SM, Delwiche CF. 2011. Dynamics of actin evolution in dinoflagellates. Mol Biol Evol 28:1469-1480.
Kim SW, Jung WH, Ryu JM, Kim JB, Jang HW, Jo YB, Jung JK, Kim JH. 2008. Identification of an alternative translation initiation site for the *Pantoea ananatis* lycopene cyclase (*crtY*) gene in *E. coli* and its evolutionary conservation. Protein Expr Purif 58:23-31.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinf 25:2286-2288.
Leterme SC, Seuront L, Edwards M. Differential contribution of diatoms and dinoflagellates to phytoplankton biomass in the NE Atlantic Ocean and the North Sea. Mar Ecol Prog Ser 312:9.
Lin S. 2011. Genomic understanding of dinoflagellates. Res Microbiol 162:551-569.
Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, Passarotti M, Kaufman S, O'Leary MA. 2015. A RESTful API for access to phylogenetic tools via the CIPRES science gateway. Evol Bioinform Online 11:43-48.
Moore RB, Obornik M, Janouskovec J, Chrudimsky T, Vancova M, Green DH, Wright SW, Davies NW, Bolch CJS, Heimann K, et al. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. Nature 451:959-963.
Morse D, Salois P, Markovic P, Hastings JW. 1995. A nuclear-encoded form II RuBisCO in dinoflagellates. Science 268:1622-1624.
Moszczynski K, Mackiewicz P, Bodyl A. 2012. Evidence for horizontal gene transfer from Bacteroidetes bacteria to dinoflagellate minicircles. Mol Biol Evol 29:887-892.
Mungpakdee S, Shinzato C, Takeuchi T, Kawashima T, Koyanagi R, Hisata K, Tanaka M, Goto H, Fujie M, Lin S, et al. 2014. Massive gene transfer and extensive RNA editing of a symbiotic dinoflagellate plastid genome. Genome Biol Evol 6:1408-1422.
Nakamura Y. 1998. Growth and grazing of a large heterotrophic dinoflagellate, Noctiluca scintillans, in laboratory cultures. J Plankton Res 20:1711-1720.
Nassoury N, Morse D. 2005. Protein targeting to the chloroplasts of photosynthetic eukaryotes: getting there is half the fun. Biochimic Biophys Acta 1743:5-19.
Nassoury N, Wang Y, Morse D. 2005. Brefeldin a inhibits circadian remodeling of chloroplast structure in the dinoflagellate gonyaulax. Traffic 6:548-561.
Nelson MJ, Dang YK, Filek E, Zhang ZD, Yu VWC, Ishida K, Green BR. 2007. Identification and transcription of transfer RNA genes in dinoflagellate plastid minicircles. Gene 392:291-298.
Nelson MJ, Green BR. 2005. Double hairpin elements and tandem repeats in the non-coding region of Adenoides eludens chloroplast gene minicircles. Gene 358:102-110.
Nisbet RER, Koumandou VL, Barbrook AC, Howe CJ. 2004. Novel plastid gene minicircles in the dinoflagellate *Amphidinium operculatum*. Gene 331:141-147.
Oudot-Le Secq MP, Grimwood J, Shapiro H, Armbrust EV, Bowler C, Green BR. 2007. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. Molecular Genet Genom 277:427-439.
Pochon X, Putnam HM, Gates RD. 2014. Multi-gene analysis of *Symbiodinium* dinoflagellates: a perspective on rarity, symbiosis, and evolution. PeerJ 2:394.

Probert I, Siano R, Poirier C, Decelle J, Biard T, Tuji A, Suzuki N, Not F. 2014. *Brandtodinium* gen. nov. and *B. nutricula* comb. nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. J Phycol 50:388-399.
Puthiyaveetil S, Kavanagh TA, Cain P, Sullivan JA, Newell CA, Gray JC, Robinson C, van der Giezen M, Rogers MB, Allen JF. 2008. The ancestral symbiont sensor kinase CSK links photosynthesis with gene expression in chloroplasts. Proc Natl Acad Sci USA 105:10061-10066.
Reynolds JM, Bruns BU, Fitt WK, Schmidt GW. 2008. Enhanced photoprotection pathways in symbiotic dinoflagellates of shallow-water corals and other cnidarians. Proc Natl Acad Sci USA 105:13674-13678.
Robbens S, Derelle E, Ferraz C, Wuyts J, Moreau H, Van de Peer Y. 2007. The complete chloroplast and mitochondrial DNA sequence of *Ostreococcus tauri*: organelle genomes of the smallest eukaryote are examples of compaction. Mol Biol Evol 24:956-968.
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539-542.
Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res 6:283-290.
Shalchian-Tabrizi K, Skanseng M, Ronquist F, Klaveness D, Bachvaroff TR, Delwiche CF, Botnen A, Tengs T, Jakobsen KS. 2006. Heterotachy processes in rhodophyte-derived secondhand plastid genes: implications for addressing the origin and evolution of dinoflagellate plastids. Mol Biol Evol 23:1504-1515.
Siano R, Montresor M, Probert I, Not F, de Vargas C. 2010. *Pelagodinium* gen. nov. and P. béii comb. nov., a dinoflagellate symbiont of planktonic foraminifera. Protist 161:385-399.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinf 30:1312-1313.
Walker JE. 2013. The ATP synthase: the understood, the uncertain and the unknown. Biochemical Society Transactions 41:1-16.
Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom Proteom Bioinf 8:77-80.
Wang YL, Morse D. 2006. Rampant polyuridylylation of plastid gene transcripts in the dinoflagellate *Lingulodinium*. Nucl Acids Rese 34:613-619.
Wisecaver JH, Hackett JD. 2011. Dinoflagellate genome evolution. Ann Rev Microbiol 65:369-387.
Xia X. 2013. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30:1720-1728.
Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Mol Phylogenet Evol 26:1-7.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496-503.
Zauner S, Greilinger D, Laatsch T, Kowallik KV, Maier UG. 2004. Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellate Ceratium horridum. FEBS Lett 577:535-538.
Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, Lin S. 2007. Spliced leader RNA trans-splicing in dinoflagellates. Proc Natl Acad Sci USA 104:4618-4623.
Zhang Z, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. Nature 400:155-159.
Ševčíková T, Horák A, Klimeš V, Zbránková V, Demir-Hilton E, Sudek S, Jenkins J, Schmutz J, Přibyl P, Fousek J, et al. 2015. Updating algal evolutionary relationships through plastid

genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? Sci Rep 5:10134.

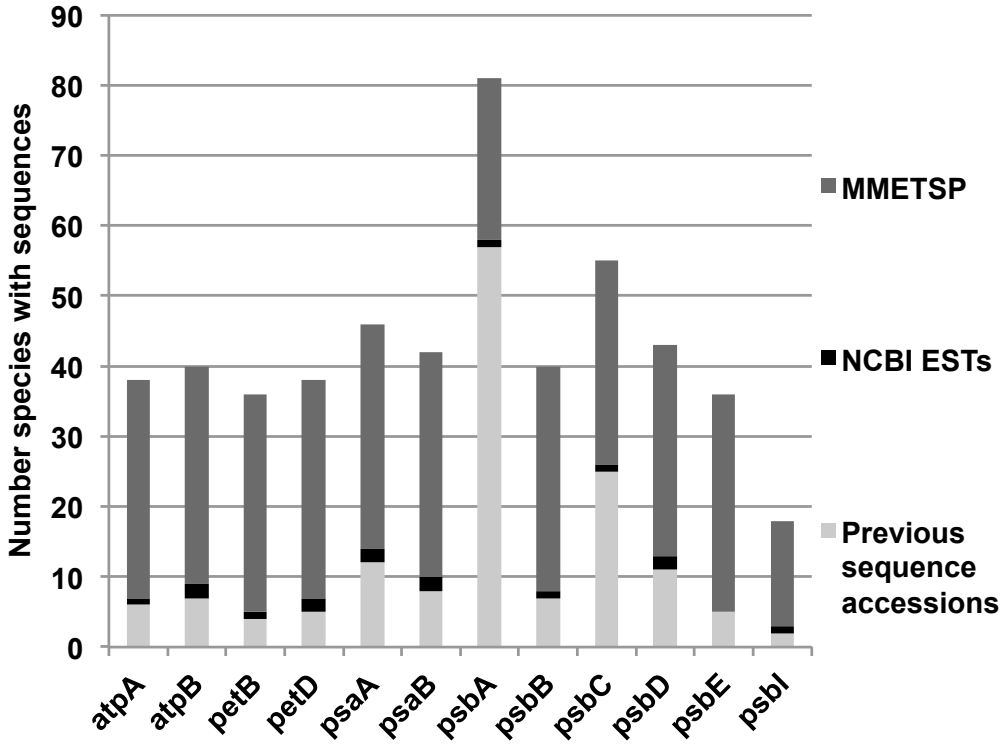


Fig. 1. New peridinin plastid coding sequences identified from transcriptome libraries. This graph shows the number of peridinin dinoflagellates for which sequences of each of the twelve protein-coding genes present in peridinin plastids have been identified, both previously, and from the data presented in this study. Sequences are only counted if they are > 200bp length (for *atpA, atpB, psaA, psaB, psbA, psbB, psbC* and *psbD*), > 100 bp (for *petB, petD, psbE*) or > 50bp (for *psbI*).

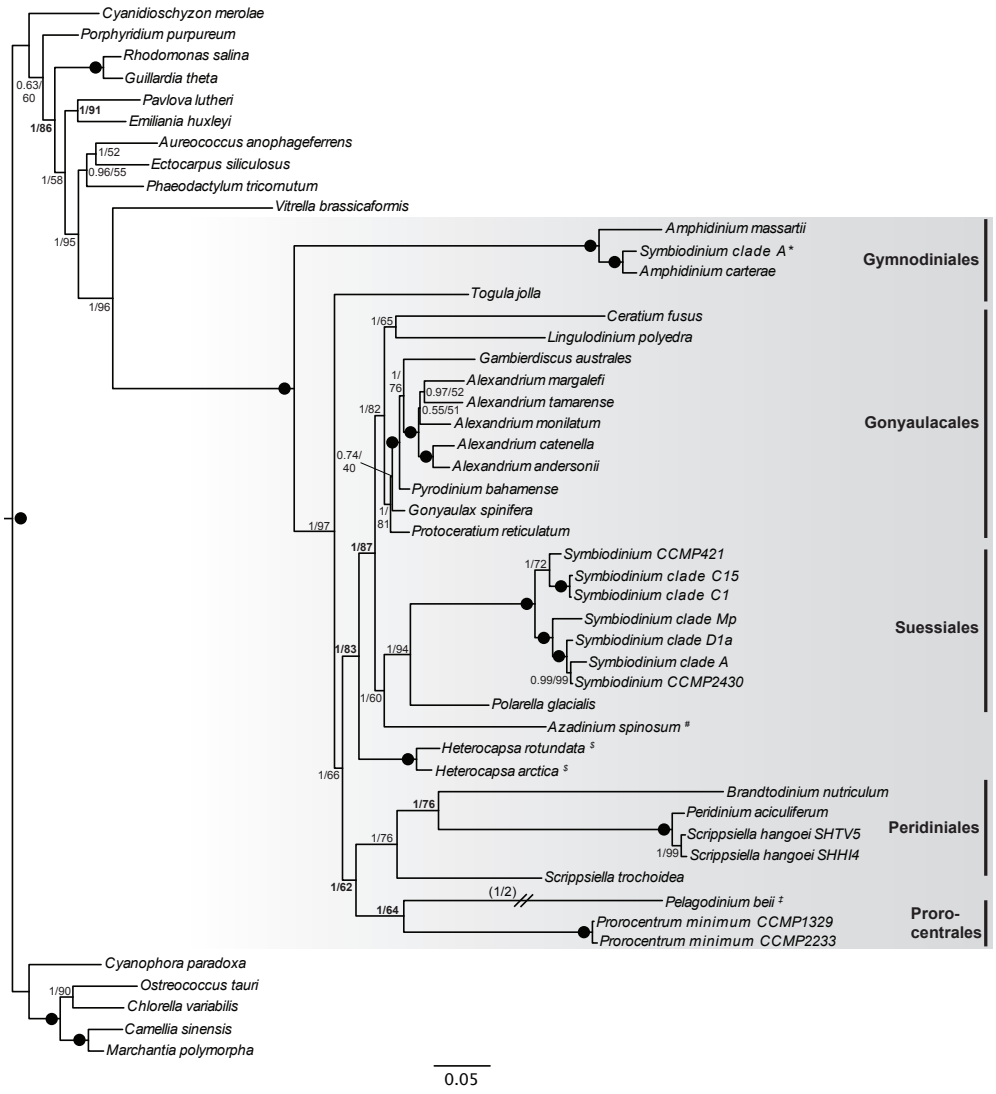


Fig. 2. Multigene protein phylogeny of peridinin dinoflagellates. This figure shows the Bayesian obtained for the 48 species x 3410 aa alignment of the twelve proteins that are plastid-encoded in peridinin dinoflagellates. Filled circles at each node indicate support with a posterior probability of 1.0 (for Bayesian inference) and 100% bootstrap values (for RAxML); elsewhere, support values for each node are given in the format (MrBayes/ RAxML). The branch leading to *Pelagodinium beii* (marked with double crossed lines) has been reduced to half its true length (true length 0.463, displayed length 0.232) to accommodate the branch within the figure. Two phylogenetically distinct populations of sequences identified from the Clade A *Symbiodinium* library are shown, one grouping with other *Symbiodinium* sequences, and the other (asterisked) within *Amphidinium*. Dinoflagellate orders (Gymnodiniales, Gonyaulacales, Peridiniales, Suessiales, and Prorocentrales) are labelled on the diagram. Species that do not resolve with other members of the same order are individually labelled, as follows: # (*Azadinium spinosum*), previously assigned to Gonyaulacales but identified here as a sister group to the Suessiales; $ (*Heterocapsa sp.*), assigned to Peridiniales but identified here as a sister group to Gonyaulacales and Suessiales; ‡ (*Pelagodinium beii*), assigned to Suessiales but identified here as a sister group to Prorocentrales.

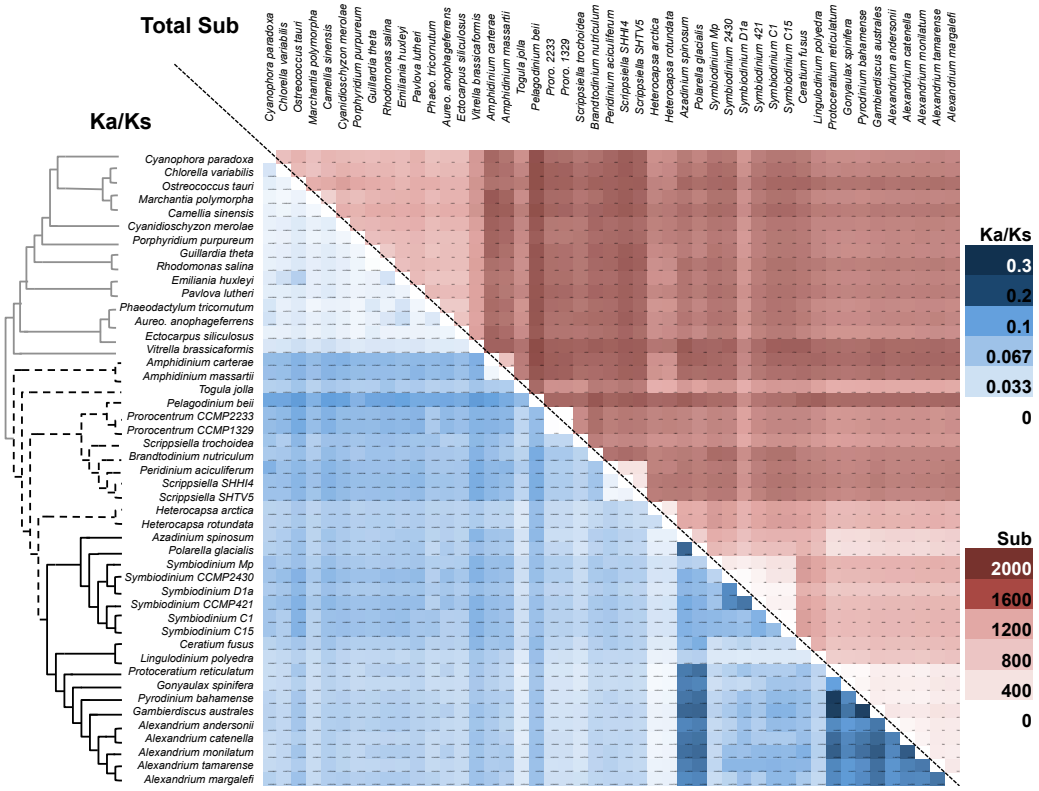
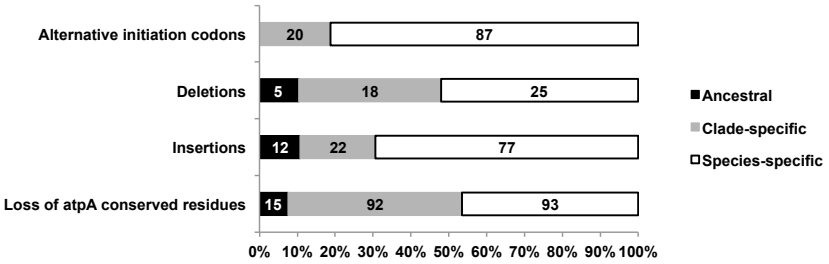


Fig. 3. Pairwise substitution ratios in peridinin dinoflagellates. This heatmap shows pairwise Ka/Ks ratios (bottom left hand sector) and total number of pairwise substitutions (top right hand sector) between every dinoflagellate and non-dinoflagellate sequence in the multigene nucleotide alignment. Sequences are shown per the multigene tree topology established in figure 2; non-dinoflagellates are shown on grey branches, basal dinoflagellates (Gymnodiniales, Peridiniales, Prorocentrales and *Pelagodinium*) with dashed branches; and Gonyaulacales and Suessiales on solid black branches. The scale factors associated with each value are shown on the right hand side of the figure.



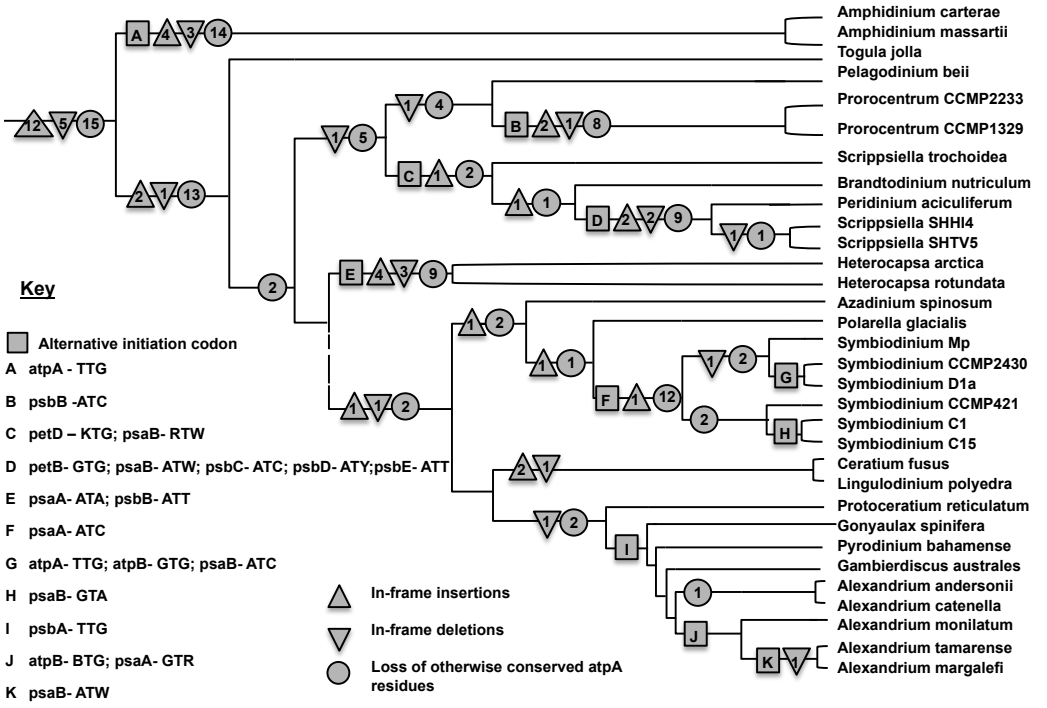
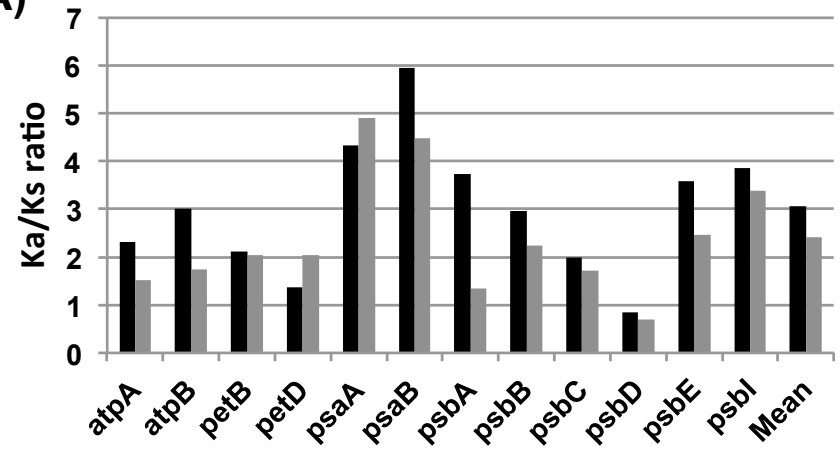


Fig. 4. Phylogenetic reconstruction of clade-specific evolutionary events in peridinin dinoflagellates. This figure shows discrete changes to peridinin plastid sequences that have occurred throughout dinoflagellate evolution. Panel A shows the number of instances of four evolutionary events (adoption of alternative translation initiation codons, in-frame insertions and deletions, and the loss of otherwise conserved residues from atpA) that originated in a common ancestor of all dinoflagellates, are confined to specific species, or that originated in the common ancestors of specific dinoflagellate clades. Panel B shows the phylogenetic origins of each of the clade-specific features, as determined by comparison to the multigene tree topology obtained in Fig. 2. Alternative initiation codons are shown with square labels, in-frame insertions and deletions with triangular labels, and discrete changes to otherwise conserved *atpA* residues with circular labels.

A)



B)

	All residues	Positive selection ancestral branch	Positive selection within dinos
Total	4121	331	517
Transmembrane domains	1008	*45***	*108***
Stromal residues	691	**70****	**156****
Luminal residues	1488	110	223
Intersubunit interfaces	472	*29**	*2***
Cofactor-binding residues	393	32	*0***

Fig. 5. Site-specific evolution in dinoflagellates. Panel A shows the Ka/Ks residues calculated for each of the twelve peridinin plastid genes. Black bars show the Ka/Ks ratios inferred for the dinoflagellate common ancestor, divided by the Ka/Ks ratios calculated for every non-dinoflagellate reference sequence within the alignments. Grey bars show the Ka/Ks ratios calculated within the dinoflagellate sequences, again divided by the Ka/Ks ratios calculated for the non-dinoflagellate sequences within the alignment. Panel B tabulates the residues with adjusted Ka/Ks ratio > 1, and Ka/Ks ratio significantly greater than corresponding non-dinoflagellate reference Ka/Ks ratio) in the common ancestor of all studied dinoflagellates or within the dinoflagellates (as above, except adjusted Ka/Ks ratio > 0.5) and their predicted functional properties. Values that are significantly higher than would be expected through random distribution of these sites are shown in bold text, and values that are significantly lower are shown in italics. Values with one asterisk are significant to $P < 0.05$, and values with two asterisks are significant to $P < 1 \times 10^{-04}$.